SUBJECT ACCESS PROJECT

SAP

IMPROVED SUBJECT RETRIEVAL FOR MONOGRAHPIC PUBLICATIONS.

Irene Wormell

SUBJECT ACCESS PROJECT - SAP

Improved subject retrieval for monographic publications.

by

IRENE WORMELL

Fil. kand, informatiker

AKADEMISK AVHANDLING

som för avläggande av filosofie doktorsexamen vid Samhällsvetenskapliga fakulteten vid universitetet i Lund kommer att offentligen försvaras å Carolinasalen onsdagen den 13 mars 1985 kl.10.

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41546224_0019

SUBJECT ACCESS PROJECT - SAP

Improved subject retrieval

for monographic publications.

by

Irene Wormell

Lund 1985

Organization
LUND UNIVERSITY
Dept. of Info and Computer Science
Sölvegatan 14 A
S-223 62 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of issue
January 1985

CODEN:

Author(s)
Irene Wormell

Sponsoring organization

Title and subtitle
Subject Access Project - SAP.
Improved subject retrieval of monographic publications.

Abstract

The thesis is based on the theoretical and practical work carried out within the range of the Subject Access Project - SAP at Lund University Library during 1977-83. Its main objective was to improve the retrieval capabilities of online library catalogs in Sweden, meeting the conditions of the "online age".

The function of indexing is studied in the relation to those radical changes which are going on in the operational environment of IR-systems.

Indexing and the documentary information transfer are interpreted as mainly communication processes relying on language and its semiotic components.

It is emphasized that the improvment of IR-systems is an area of research which draws its substance, method and technique from a variety of contributory disciplines.

To gain a deeper insight into the various functions of bibliographic work, the accessibility of graphic records is elucidated as a tripartite conception, such as, physical, bibliographic and intellectual accessibility.

Key words
Indexes, Text analysis, Cognitive view, Syntactics, Semantics, Pragmatics, Notification Hexagon

Classification system and/or index terms (if any)
UDK 025.3 025.4 087.7

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN

Recipient's notes

Number of pages
147

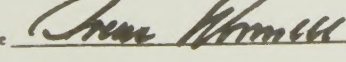
Price
95:- SEK

Security classification

DOKUMENTDATABLAD enl SIS 61 40 21

Distribution by (name and address) Irene Wormell
Gl. Kongevej 31.I DK-1610 Copenhagen Denmark

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature  Date January 11, 1985

CONTENTS

INTRODUCTION

1 SUBJECT ACCESS PROJECT - SAP 1-9
1.1 The background of the project
1.1.1 Inadequencies of present library catalogs
1.1.2 Changed conditions in library catalog design
1.1.3 The origin of SAP
1.2 Research areas and problems involved in the project

2 THE INTELLECTUAL ACCESSIBILITY OF LIBRARY COLLECTIONS 10-15
2.1 Bibliography as an instrument of the total communication process throughout society
2.2 The tripartite conception of accessibility

3 SEMIOTICS AND THE DOCUMENTARY INFORMATION TRANSFER 16-24
3.1 Syntactics
3.2 Semantics
3.3 Pragmatics
3.4 Indexing as a communication process

4 THE COGNITIVE VIEWPOINT 25-30
4.1 The cognitive stage in thinking on information processing
4.2 Cognitive sciences

5 ELEMENTS OF IR-SYSTEMS 31-40
5.1 The Notification hexagon
5.2 The analysis of SAP-indexing by the Notification hexagon

6 EXPERIENCES AND FINDINGS OF WORK WITH SAP 41-48

7 FINAL REMARKS AND SUGGESTIONS FOR FURTHER WORK 49-51

LIST OF REFERENCES

The essay is a summary of the following six papers:

1. "Subject Access Project - The use of Book Indexes for Subject Retrieval System in Libraries."
   Int.Forum Inf. Doc., 6(1981)4, pp.24-28.

   (Paper presented at the 39th FID Congress, Copenhagen, August 1980).

2. "SAP - a new way to produce subject description of books."
   Journal of Information Science, 3(1981) pp.39-43.

3. "Subject Access Project, SAP. En ny test-databas i ESA-QUEST."
   Tidskrift för Dokumentation, 37(1981)4, pp.75-81.

4. "SOU-data base. Retrieval of factual data."
   Menneske-System-Maskin. Proceedings of NORD IoD 5.
   Norsk Dokumentasjonsgruppe, Oslo, 1982, pp.317-325.

   (Paper presented at the 5th Nordic Conference on Information and Documentation, Trondheim, June 1982).

5. "Cognitive aspects in natural language and free-text searching".
   Social Science Information Studies, 4(1984)4, pp.131-141.

   (Paper presented at the NORDINFO Seminar on the Psychological Aspects of Information Searching, Copenhagen, November 1983).

6. "Indexes and text analysis in automated IR-systems".
   The Royal School of Librarianship, Copenhagen, 1984.

INTRODUCTION

The present work describes an approach to development of modern library catalogs, meeting the conditions of the "online age". The thesis is based on my theoretical and practical work carried out within the range of the **Subject Access Project-SAP** at Lund University Library during 1977-83. (Except for a period of 12 months when in 1980-81 the author was working as a research fellow at ESA-IRA, the European Space Agency Information Retrieval Services in Frascati, Italy).

SAP-project was financed by SINFDOK/DFI and its main objective was to improve the retrieval capabilities of online library catalogs in Sweden.

SAP has resulted in a new indexing method for subject description of monographic publications, which enhances subject access in library catalogs. It gave purposeful suggestions for the realization of the multi-access points bibliographic records. Knowing the importance of time and cost constraints in bibliographic work, SAP meets the present economic conditions for the development of online library catalogs.

Through the application of the SAP-indexing method two data bases have been developed, which can be considered as the practical results of the project. The first has been a test file containing Swedish and English monographs. For the description of the detailed subject content of the books, selected parts of the table of contents and the back-of-the-book-indexes have been used. The SAP-file has been functioning as an online catalog, offering highly specific subject searches via free-text search teqnique.

The second file represented further development of SAP and a new application-area to make all tables and graphs included in the books searchable. To test this new idea the Swedish Government Committee Reports, SOU-series, were chosen as a basis. The SOU-data base contained about 1000 reports published 1973-82, and it represented an unique source for factual information.

A convincing result of the project, is the fact that SOU-data base has been taken over by DAFA and has been in commercial use since May 1983. It is publicly available as an integral part of the central Swedish information system for legislation and public administration (RÄTTSDATA).

With the exeption of some assistance for input at the creation of the data bases, I am entirely responsible for the carrying out of the project.

The results of the project have been reported to the library world and publicized in three comprehensive reports and several articles in national and international journals.

The articles and conference papers included in the thesis represent a practical description of the different phaces of the project, while the study and the summary aim to give an analysis of the theoretical problems encountered and handled during the SAP-project. It should be noted that personal references in these publications are intended to designate both sexes, but for the reasons of economy the shorter form 'he' is used.

The work presented has been initiated by the fact that information is abundantly generated and systematically stored in libraries but not yet efficiently used. In this connection SAP is to be considered as a "fragment of a larger pattern". It attempts to raise the efficiency of information systems which yield only references to documents and do not assure the reader to find the information in a context to which he can relate.

The thesis aims to emphasize the fact that the improvment of IR-systems requires the utilization of theories and ideas scattered among different scientific disciplines: Since the relation between "signifiant" and "signifé" (as it is defined by Ferdinand de Saussure) represents one of the basic concepts in information retrieval, all discussions about the representation of document contents in autometed IR-systems are strongly related to the field of semiotics. Furthermore, the problem of documentary information transfer involves several aspects of communication as well as the understanding of language. By the recognition that information and knowledge are entities which exist in cognitive spaces, the searching and retrieval of documents must be seen as mainly cognitive processes relying on language and its semiotic components.

To gain a deeper insight into the various functions of bibliographic work, the accessibiblity of library collections is elucidated as a tripartite conception, such as, physical, bibliographic and intellectual accessibility.

Hopefully, this analysis will cover, to some extent, the opinions and theories about library catalogs and will be a useful basis for further discussions, to improve the art of bibliographic access in Sweden.

My research work has been carried out under the supervision of Björn Tell, the former Director of Libraries at Lund University, to whom I wish to express my gratitude for the stimulation and encouragement which he has given to me in my professional activities. My thanks are further due to Professor Kjell Samuelson who conducted my studies in the field of Informatics with General Systems and Cybernetics at the Royal Institute of Technology and the University of Stockholm, and to all my friends and colleagues who, in many different ways, have given me help and support.

Lund-Frascati 1984

1. SUBJECT ACCESS PROJECT - SAP

> "You will seek not a near, but a distant, objective, and you will not be satisfied with what you may have done. All that you may achieve or discover you will regard as a fragment of a larger pattern, which from his separate approach every true scholar is striving to descry".
>
> James G. Miller

1.1 The background of the project

1.1.1 Inadequacies of present library catalogs

When considering the question of what function catalogs are supposed to fulfil, and to what extent they actually do it, I want to refer to Cutter's classical statement (1904):

> "...**the 'object' of a catalog** is to show what the library has by a given author, or on a given subject, or in a given kind of literature. (Observe the formulation which did not say that the catalog was to show what books a library had, but simply, **what** it had). Thus the main function of the catalog is to enable a person to find a document of which either the author or the title or the subject is known." (The italics are added by the author)

Although most people agree that the catalog is supposed to enable the user to find information, which will help solve or at least clarify his problem, some parts of the "object" of the catalog have been unduly neglected, and most of the existing catalogs do not properly fulfil this function. Conventionally the catalog is supposed to be an efficient instrument only for ascertaining known items.

Accessing information can be quite a mysterious and irrational process to non-librarians and library catalogs cause embarrassment sometimes even for professional users.

Many inadequacies of present library catalogs are rooted in the fact that they were primarly designed for physical description and identification; and not for subject searching. The traditional choice of basic catalog units has the consequence that **enormous numbers of texts are not locatable through the catalog.** Catalogs do not aim to show all the versatile and distinguishable contents of a book and therefore a lot of

valuable information remains "hidden" for the searcher. Our investment in books is partially wasted and underused if access is available only to those who come to the library to search for known items. Thus, the catalog should help one to find copies of texts without knowing their published appearance. In short, it is supposed to be a retrieval instrument for the book-finder as well as for the text-finder.

Users of online catalogs in the last decade were voicing their needs more and more for enhanced subject access. But as long as the library catalog remained in card or microfilm format, it was easy to ignore this demand - the linear structure and the physical format of the catalog presented a barrier to adequate subject description.

This does not mean that discussions about cataloging rules and investigations into the process of subject description and retrieving have not taken place. However, these activities have not to any serious degree influenced or changed the premises of subject cataloging in libraries. Ignoring the **demand for flexible and suitable retrieval tools** for the various user groups, librarians still spend much more time and energy on the traditional bibliographic description than on content analysis (White 1981) (Hildreth 1983) (Markey 1984) (Atherton-Cochrane 1984).

In spite of the importance of **time and cost constraints** in bibliographic work, librarians have, strangely enough, not been interested in examining and revalidating the old premises of subject cataloging; they are stuck with content description in terms of subject matter and form (in a "l'art pour l'art" manner).

All the considerable investment of time, money and effort in recent years, in the creation of improved bibliographic records (both at the international and national level) have had no real impact on the librarians' routines for subject cataloging. Although we have a whole range of library networks operating today all over the world, and a great portion of existing bibliographic records are in machine-readable form, none of them have adequate subject descriptions of the books described therein.

The utility of existing library catalogs is restricted by the fact that they provide very little information about the varied and extensive contents of the books.

The use of computers in the library has until now, mainly implied only a transfer of the old manual routines to automated processes - without a real extension of retrieval capabilities for topical information.

1.1.2 Changed conditions in catalog design

Traditionally the question of catalog access has always been equivalent to the decision about entries: main entry and added entries. The tedious old notion of main entry in manual catalogs, implied that an item appears once at a certain place in the catalog (either under the name of the author, or the title, or the name of the series) giving the full descriptive information about the document. The other appearances of the same item in the catalog were considered as added entries, giving only partial information.

The traditional shallow subject description in the manual catalogs is related to the fact that in the design of entries one was bound to the information which could be put on a 12 x 8,5 cm card.

However, in modern catalogs which are created in machine-readable form, the main entry does not fulfil any purpose. Generally **the notion of entry is transferred into the question of permissible search requests** - the amounts and formats of information we will allow the user in order to make the item appear. Although the record for an item is physically stored in one place in the system, it is not any more the item's main entry but rather the source of all possible entries - which can be many different kinds of search keys, search statements, clues and combinations of clues. **This practical difference in the notion of entry has radically changed the old question of how to make bibliographic records accessible.**

Utilizing the **capacity of computers to manipulate large files of information,** we can now enter much more information about a book's subject into a file of machine-readable records, thus increasing the number of different elements of information that might be used for retrieval. It means that (in practice) we can now deepen the shallow subject indexing in our catalogs.

The **combinatorial capacity of computers** is the other fact which is considerably increasing our search possibilities, compared with the accessibility of main and added entries in the manual systems. By logical and positional operators, we can combine clues and partial bibliographic information in many sophisticated ways, using different parts of the catalog record. This flexibility in the formulation of search requests, is a very important feature of modern IR-systems: it extends for the user possibility to gain access to information from many different points and partly eliminates the use of meta-languages in the interaction with the system.

When searching factual information in SOU-data base, e.g., an effective search strategy is to use the name of the department for the first rough grouping of Committee Reports and from this

level, conduct the search further in the vast amount of tables and graphs. Grouping by publishing year or the name of the Committee, are also usable bases to start the search with.

1.1.3 The origin of SAP

In the late 1970s, further development of existing online catalogs was a controversial question in Sweden: there had been many different opinions and ideas about activities concerning library catalogs.

The common concept, that the primary function of the catalog is to find and identify known bibliographic items, has strongly influenced the design of the two Swedish computer-based bibliographic systems - LIBRIS and BUMS (the first adherent to research libraries, the second to public libraries). This belief, however, has been dominant in bibliographic work, not only in Sweden but all over the world (Brookes 1981) (Atherton-Cochrane 1983).

There were only a few professionals who argued for devoting more resources to the improvement of topical subject descriptions in library catalogs. Developments in computer technology have greatly aided this demand and some researchers took up the challenge to see how it was possible to convert the traditional **unifunctional** catalogs into **multifunctional** online catalogs - where both known-items and topical subject searches can be successfully completed.

In 1977, when I was looking for research objects, the recently finished **Subject Access Project (SAP)** at Syracuse University, School of Information Studies made a deep impression on me (Atherton 1978). Since both my tutor Björn Tell and myself were convinced that it was necessary also in Sweden to find new and creative ideas for catalog development, it was decided that the project at Lund University Library was to start on the bases of the **SAP-indexing method**, developed at Syracuse.

The project at Syracuse has introduced the acronym SAP to the library world as a new indexing method. Since the project in Lund was a Swedish version of that it was advantageous to keep this acronym. Therefore, great effort has been made find a corresponding Swedish name for the project. The result has been the following somewhat intricate phrase "Slagords-automatiseringsprojekt" (Wormell 1978).

When it is necessary to indicate that SAP refers to the work which has been carried out at Lund University Library, the acronym will be used here in a modified way: **SAP-S** (Sweden).

The main objective of the project in Lund was to create augmented subject description for existing MARC-like catalog records in Sweden, by selecting index terms from the table of contents and the back-of-the-book-indexes.

The selection of index terms was governed by easy, quantative rules, described in detail in the **"SAP-indexing manual"** (Settel 1977). The augmentation of subject description was running up to an average of 300 words per title, or about 30 entries per book. An entry was a group of **terms in a contextual phrase** such as a chapter heading or indexing phrase.

The promising results of SAP at Syracuse have represented a great challenge for me in trying to apply innovative ideas in the creation of online catalog records in Sweden. The experiences and the results of my work are presented in Chapter 6.

1.2 Research areas and problems involved in the project.

I aim to describe here the nature of the research problems involved in SAP-S and the reasons they were selected.

1. The suggestion that **the inherent attributes of the documents can be used as the source for their representation,-** has been the underlying theoretical framework for my research activities.

   With SAP-S we wanted to test the possible usefulness of the content tables, back-of-the-book-indexes, captions of tables and diagrams for creating **detailed subject description** for the monographic publications in Swedish library catalogs.

   To use the inherent attributes of documents to indicate their subject content, in itself does not represent an original idea: there have been some other people before thinking in similar terms, e.g., H.P. Luhn's approach about 20 years ago, in the development of the KWIC Title Indexes, was based on a similar idea (See Wormell 1984a, pp. 9-14). However, the way in which SAP-indexing has applied this idea in an online environment, does mean a new and innovative method of indexing which has never been tested before.

2. SAP belongs to the large group of research activities aimed to test whether **the conventional technique of indexing documents could be substituted by automated or semi-automated indexing routines.**

It means that one partly tries to replace the exacting and costly process of traditional indexing by automated processes, where the conditions for successful search more and more rest on the computer's capacity to process text in natural language form.

For the last two decades, the research field in information science has been dominated by the problems discovered by these approaches. This trend in the development of IR-systems led to cheaply established and useful data bases, but it did not offer any realisable methods for a deeper analysis of the information content of documents (See Wormell 1984a, Chapt.2).

SAP-indexing can be labeled as a semi-automated routine where the selection of index terms is guided by quantitative rules which are easy to apply.

This technique offers a simple and straightforward way of describing or summing-up the content of the books. (With or without human intervention) it can produce representations for the documents simultaneously, both as physical objects as well as potential sources of information. Instead of relying on sophisticated counting routines in measuring information, the "successful matching" in subject searching is primarly based here on the supposed **nearness of the authors' and the users' vocabularies.**

The fact that **SAP-indexing is mediating the documents via the authors' own vocabulary,** makes it unique among other methods for deep indexing. Having experience from many different IR-systems I have found this characteristic extremly valuable and it has represented the cornerstone of my own approach to enhance subject access in Swedish library catalogs.

3. In my research activities I have tried to pay attention to the human user of IR-systems. Therefore SAP was also an approach towards bring catalogs closer their users. The main idea behind this approach was that **catalog records, augmented by index terms and phrases in natural language form can facilitate the use of catalogs** (Miller 1978, Hypothesis 3.3.3.2-16, pp 97-98).

The use of present library catalogs is partly limited by the fact that classification codes and subject headings (selected from a controlled vocabulary) seem to be efficient retrieval tools only in the hands of professional bibliographers and/or subject specialists. The use of different meta-languages can render difficulties in certain user environments. Thus, the various access points in

natural language form, possibly combined with classification codes and index terms, **offer the user great freedom and flexibility to use whatever information he wants to start his search with** - in trying to find a book or locate particular items in the library's collection.

4. Having knowledge of the **General Systems Methodology & Synthesis and Living Systems Theory,** I have applied a multidisciplinary systems approach to the study of indexing in online library catalogs.

Synthesizing systemic findings from specialized fields such as, information science, science of science, computational linguistics, social and cognitive sciences etc., I have tried to put bibliographic work in a wider perspective.

As bibliographic organisation means a pattern of effective arrangement of the records of human communication, it implies the transfer between physical documents, their representation, and the intellectual content of documents (See Chapter 2).

In my study of the functions of indexing, I have been guided by the general theory of living systems, where information transfer is discussed in detail at the different levels of the biological and social systems. James G. Miller's great work **"Living Systems"** (1978) is a unique general checklist offering detailed analysis of the major aspects and characteristics of the matter-energy and information processing subsystems, encountered at all the seven levels of the systems he distinguishes.

5. In thinking about efficient methods of coding in systems dealing with documentary information transfer, Miller's cross-level hypothesis concerning living systems have served me as intellectual guides: **Encoding** and **Decoding,** as information processing subsystems, have been studied at the level of Organism, Group, Organization and Society (See Section 3.4) **indexing** has been understood **as a communication process** mediating information from the author to the reader. Thus, the present work aims to emphasize that, if we want to be able to organize the "products" of human thinking in such a way that the social utility of document records can be maximized for the benefit of mankind, we need to understand how people think and interact with IR-systems (See Wormell 1984a, pp.20-24).

The questions of text and knowledge representation led me on to problems dealing with the structure and function of mental representation. Studies in the allied fields of

"cognitive sciences" helped me to gain a deeper understanding of indexing as part of human communication (See Chapter 3 and 4). In my research I have tried to draw attention to how much the individual **view of the world** influences the interaction between the user-intermediary-system. Since the development of IR-systems is closely related to the problem of human information processing, I am convinced that, only by an **integrative approach,** can we achieve veritable results in this research area.

6. The development of SAP-S has been also guided by the fact that, all progress in catalog development has to imply **suggestions which are economically feasable.** One has to pay attention to the fact that labor costs for the intellectual work of the indexer are continually rising, while the costs for automatic processing are predicted to decrease.

7. It should also be mentioned that SAP-S has been an **approach to encourage the use of free-text technique in libraries.** In the last two decades the computer based abstracting and indexing services, offering advanced search facilities, have been widely used at documentation centers. Until now libraries did not take advantage of the improved subject search capabilities, such as the use of keywords, truncation, masking, logical and positional operators etc.

8. SAP has contributed to discovering and emphasizing the changes that occur in subject access process, as users move from card catalogs and printed indexes to online catalogs and bibliographic systems, where subject access is possible through controlled vocabulary as well as free-text keyword searching, with or without the help of intermediaries.

   To put SAP-S in a **historical perspective,** it should be pointed out that in the late 1970s these changes had still not been observed by the library world with that sharpness and intensity which is characteristic for the 1980s. Particularly in Sweden the author was rather alone in working on the utilization of free-text search technique in library catalogs.

   Beside the increase of all the new data bases - offering sophisticated topical searches in journal articles - the problem of subject access for books in library catalogs had been concealed for a long time.

In the 1970s most librarians considered PRECIS as the "next step forward" in catalog development. PREserved Context Index Systems, developed by the staff of the British National Bibliography, produced computer aided printed bibliographies on the basis of the semantic interpretation of transformational theory (Austin 1974) (Hedberg 1976).

Thus, SAP-S (offering free-text serches in library catalogs) represented at that time a new radical thinking in catalog development. This is also partly true today. However, catalog research is now under way.

In the last two years there have been published e.g., a great variety of studies on **"Online Public Access Systems" - OPAC** as the result of a multifacetted research project in the US, sponsored by the Council on Library Resources, Inc. Extensive research activities are going on at OCLC (Online Computer Library Center, Dublin, OH) and several other places in the US. In 1980 a Centre for Catalogue Research was built at Bath University and similarly in Sweden, at the University of Linköping; where an ambitious research program was started in 1982 by the Library Laboratory (LIBLAB). The problem of subject searches has been brought up for discussion also in Denmark and has initiated some interesting experiments there (Lau and Winkel 1983) (Online emnesøgning 1983) (Emnedata i online-alderen 1984).

2. THE INTELLECTUAL ACCESSIBILITY OF LIBRARY COLLECTIONS

Remembering the mosaic script on the inside wall of the entrance to the Library of Congress in Washington which says that **books are the world's greatest treasure,** one is motivated to reflect on how limited our possibilities are to get hold of this treasure.

The idea of SAP is rooted in the fact that during the 1970s the number of computer-based abstracting and indexing services, providing detailed subject access to journal articles, increased remarkably, whilst little has been done to improve the retrieval capabilities for the user of monographs.

In spite of the considerable investment of time, money and effort e.g., in the development of a whole range of library networks or in the creation of new editions of the Dewey Classification scheme and the Bliss Classification scheme - still, present library catalogs have a very limited value.

Every available test result indicates that the use of any of the standard library classification systems does not yield sufficient results in a retrieval environment. Library catalogs work well when an author or title is known; however, they are less efficient for retrieving material by subject. It is well known how defective existing classification systems are when it comes to more specific and detailed subject retrieval. Our investment in books is partially wasted and underused if access is available only to those who come to the library catalog to search for known items.

The utility of bibliographic data bases using traditional retrieval tools, whether an individual library's catalog or the full set of MARC-records in a very large system, is restricted by the fact that the individual records provide very little information about the content of the books.

According to **Popper's philosophical model of the three worlds** (Popper 1973), - where everything in existence and in experiences subsumed in one or the other of these three worlds - the objective knowledge of World 3 (the man-made world of culture) is encoded in the various physical objects of World 1.

Consequently a book has a double representation in this model: it can be perceived as an artifact, belonging to World 1, but also as a media, bearing knowledge in an objective sense in World 3. **Therefore library catalogs should provide access not only to the physical properties of books but also to the man-made intellectual work contained in them.** (Wormell 1984a, p.20).

But today's library catalogs are designed primarily for physical description and identification and not for subject access. The rich contents of a book, information full of nuances, is coded by one or more class labels or subject headings selected from pre-coordinated, controlled vocabulary lists.

The paucity of content representation of a work of perhaps several hundred pages is there for all to see. It contrasts clearly with the content representation present in many other bibliographic data bases where a work as short as a few pages in a scientific journal may be represented by several descriptors and possibly an abstract. McClure, who in 1976 provided a quantification of this contrast, found that on average 4.5 subject headings were used to represent the content of a journal article of on average 7 pages in CIJE (Current Index to Journals in Education) that on average 4.9 subject headings were used to represent the content of reports of average length 86 pages in RIE (Research in Education) whilst in the National Union Catalog the average number of subject headings used to represent the content of books is 1.3. There can be little doubt that the average page length of the books was greater than 86.

Though in computerized systems bibliographic records became more detailed, and the number of subject descriptions may be increased, yet they do not actually extend the retrieval capabilities of monographic publications. In explanation we can state that traditional manual methods for bibliographic description have been carried over into the computer age, without the realization of new potentialities for creating subject access to the contents of books.

Few librarians realize that the main drawbacks of existing subject catalogs are the arbitrary or artificial decisions which the indexer must make at the input stage when he makes his initial assignment of subject headings or classification codes to documents. These labels represent one person's opinion of the most important subject of the book. Moreover, the semantics of the headings may change with time and the concepts underlying the indexer's choice may themselves change.

Consequently the library user's knowledge of his subject can markedly differ from that subject knowledge which has once been coded by the cataloger - and therefore he may never find the information needed.

Considering this cognitive problem in information retrieval Norbert Wiener's remark from 1960 seems to be very true: "Labels are only good for preserve jars" (Doyle 1975,p.152).

2.1 Bibliography as an instrument of the communication process throughout society.

Historically the role of bibliography in the ordering and listing of recorded knowledge has been understood in various ways. Some people see it as a bare listing of documents arranged in some pattern, while others consider it to be an instrument of communicating knowledge. Theoretical studies of bibliography and bibliographic control have been made by many, and they reflect several different views. There were a few in the profession, who already three decades ago, were urging a broader perspective to the philosophy of bibliography than was traditio-nally conceived among librarians. I refer here to Jesse Shera and Margaret Egan (1952), two of the eager spokesmen of the **macrocosmic** view of bibliography - as one of the instrumentalities in the total social process of communication. The representatives of the **microcosmic view,** on the other hand argue that "a bibliography, in the correct sense of the term, is an uncritical listing of all books and materials within the coverage of the bibliography" (Stokes 1965) or that "bibliography has nothing whatever to do with the subject or literary content of the book" (Greg 1945).

Considering bibliography in the hierarchy of relations as a self-regulatory device, as a "holon" in its own right in the IR-system, we get the correspondence to the limited, microcosmic view of retrieval devices where each bibliography is regarded as a separate tool, fashioned to meet the specific needs of a limited number of persons with more or less common interests, each separate bibliography having little or no acknowledged relationship to any other. But at the same time they are subordinated as "parts" to the higher centers in the hierarchy; they form integrated structures for ordering and chartering recorded knowledge in human communication processes.

In a broader context the purpose of bibliographic work is to maximise the social utility of document records for the benefit of mankind. Taking the premise that information science is concerned with the effective transfer of desired information from human generator to human user, **the role of bibliography is understood as an instrument of communication and dissemination of knowledge.**

Putting information science and the phenomenon of information in their historical development I want to point out that in the last decade they became more and more objects for studies as their relevance for society has completely changed (Belkin-Robertson 1976), (Belkin 1977). There are theorists who go as far as to say: "Nowadays the problem of transmitting knowledge to those who need it is a social responsibility, and this social responsibility seems to be the real background of information science" (Wersig and Neveling 1976).

Shera, declaring his principles on "social epistemology" already in 1968, contends that librarians as the organizer and promoters of knowledge and as mediators between man and his communication records, must have an understanding of the sociology of knowledge.

To master the complex problem of access to recorded information requires of the bibliographers a deep comprehension of the nature of knowledge itself and an appreciation of its role in society.

2.2 The tripartite conception of accessibility

The **accessibility of library collections is viewed here as a tripartite conception** - physical, bibliographic and intellectual. Subject bibliographies and catalogs are, traditionally, our intellectual channels to the recorded knowledge of mankind. Therefore, the central organisational and intellectual problem in the design of a bibliographic system is how to gain **both intellectual and physical access** to the library collections.

The author's assumption is that access to library records is not synonymous with access to the contents of document items represented by the records; one can gain bibliographic access to records without necessarily gaining intellectual access to the information contained in the document items themselves. **Bibliographic access** is thus separated from intellectual access.

It should be noted, however, that the problem of the intellectual accessibility of books extends much further than the province of librarianship; the user's ability to understand the contents of a book in the library may have nothing to do with the intellectual accessibility of the book. It might be due either to the individual's power of comprehension or to the choice of reading material.

In view of the tradition within librarianship, to consider classification schemes as the main instrument of bibliographic organization, in most libraries the accessibility of collections is supplied only physically or, at best, bibliographically.

Most of the existing bibliographic systems do not really present an abundance of intellectual problems in their design and they are experienced by their users as far removed from the realities of human communication processes. That most science bibliographers focus their professional activities on documents as physical entity rather than recorded information, is partly

due to insufficient education in communication, information science, computer science and in the behavioral sciences.

According to the above mentioned conception, the aim of classification schemes in libraries is to provide **physical access** to their collections; to place books into an order which enables the user to find, in a certain place, books treating the same subject. Therefore, classification is best regarded as a function designed to physically bring together material likely to be required at the same time.

The limitations of the existing classification schemes as retrieval devices, arises from the fact that these systems are based upon the book as a physical entity. However, **a book represents not just a physical entity** (Popper 1973). By its inherent nature, it is a composite, poly-dimensional intellectual product which can hardly be put into a limited physical mono-dimensional system occupying only one linear position.

This gives rise to the question, is it on the whole possible to produce bibliographies for the discovery and presentation of information on the contents of documents?

> "What good is a bibliography for intellectual access? Why do people look through subject bibliographies, periodical indexes and catalogs when they could look at books and journals directly? When we create a bibliography we select certain pieces of information about and from each item to be described and array each item and its information in a list. That's what bibliographies come to, just lists of other items, usually with only very minimal bit of information about each of the items. Why do people in and out of the library field go on endlessly arranging, selecting, and rearranging these little snippets of information about sources of information?" (Bates 1976).

These questions on the "raison d'etre" of subject bibliography - as a channel of **intellectual access** to recorded knowledge - are leading the traditional thinking in librarianship into entirely new paths. Fresh and provocative ideas provide here, a further theoretical framework for bibliographic organization and control.

According to Kochen (1974) the library can potentially serve as a "growing encyclopedia system" for summarizing and organizing knowledge. They share with museums and monuments the duty of our cultural heritage. Because of their responsibilities for the custody of archival materials, librarians are understandably reluctant to consider revolutionary innovations in their art and they seldom welcome innovations or reforms from outside the world of librarianship. But the rapid changes in literature and in the demand for library services, accompanied by the equally rapid changes in technology, constitute an outside

pressure on libraries. The slowness with which these institutions usually respond to the demand for modern services mainly depends on the lack of a more responsive adaptation on the part of librarians.

It is important to point out that **classification systems increase our understanding in an extremely powerful manner.** By insisting on associations within classes of items they lead us, as it were, along paths in the undergrowth of knowledge. But, because it is easier to follow old paths than to make new ones, they sometimes mislead us and we fail to see new associations that would help us on our way.

A new consciousnes about the breadth of the problem of accessibility is reflected by the fact that nowadays several prominent scholars attest that **classification and indexing are regarded as identical with the organization of knowledge** and consequently they can be considered a subfield of the science of science (Dahlberg 1980) (Garfield 1979).

Many librarians have been aware of the limitations of classification schemes, or of the difficulties in applying them in practice. The usefulness of conventional classification has been debated by many over the years, and it is also a crucial point in present library research. I refer here only to one of the earliest researchers, Grace Kelly, who in a doctoral dissertation published in 1937, stated that library classification reveals only a very small proportion of the total library resources available in any given subject (Kelly 1937). She described the basic elements affecting the usefulness of conventional classification and summed up what many others have repeated in recent years. She pointed out that knowledge and language are dynamic processes, and that the accurate labeling of the subject contents of a book at one point in time may later be imprecise and out of context.

The fact that the need, the interest, the material - everything - will change with time, makes **the static perfection of classification schemes** impossible. As Derek J. de Solla Price expressed it during his visit to Stockholm in June of 1982:

"Last year's index is no good for this year's science".

3. SEMIOTICS AND THE DOCUMENTARY INFORMATION TRANSFER

Since language is the major vechicle for storage, retrieval and transfer of information, the problem of indexing and document representation is chiefly concerned with linguistically encoded information (Zunde 1983). This chapter is aimed at presenting the semiotical problems of information transfer and language understanding, taking into consideration a broad range of communication aspects. In semiotics the elements of human language are **signs**, the usage of which is determined by syntactical, semantical and pragmatical rules..

Early in the history of traditional grammar the question arose of the relationship between words and the "things" they referred to or "signified". It became customary to distinguish between the meaning of a word and the "thing" which was "named" by it. The following well known diagram illustrates the relationship between - name, meaning, referent (Ogden and Richard 1949).

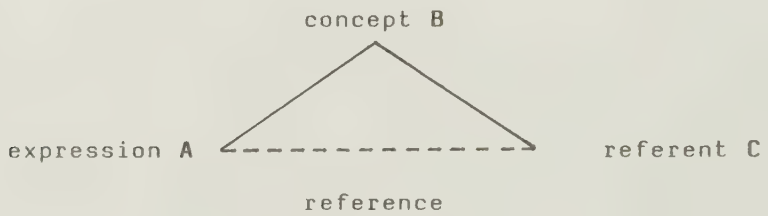


Fig. 3.1 The Semiotic Triad

Fig. 3.1 illustrates how the relation between Expression **(A)** and its "meaning" are divided into Concept **(B)** and Referent **(C)**. Referent indicates here the object in the non-verbal world, and the broken line **A-C** denotes the word and its antecedent beyond the language. The relation between Concept and Referent, line **B-C**, has been discussed since the age of scholasticism. One summarised definition is that it is that part of the "meaning" which goes beyond the referent. It is some form of "general meaning" which, depending on one's philosophical view, is the metaphysical idea of the object or the individual's mental conception of a "typical" object.

During the shortlived and intensive period of research activities in automatic indexing in the 1960s, the potential benefits of syntactic and semantic analysis in information retrieval were recognised and natural language processing came to be studied at the higher level of semiotics (Wormell 1984a, Section 2.1-3).

However there was little contact between linguists and information scientists and **the expected contributions of linguistics to** the development of IR-systems failed to keep pace with the rapid technical development in the field.

> "The intellectual task of developing algorithms that will allow the machine to detect meaning through the recognition of signs and sign sequences, the way a human being can, has not yet been accomplished. Researchers in machine indexing have not been able to deal with connotative meaning, a concept that relates to the pragmatic level of semiosis, i.e. the study of the relation of signs to their interpreters". (Artandi 1976 p.236).

In spite of the fact that it has until now been difficult to apply linguistic techniques to retrieval objectives, most of the scientists (prominent in the field today) are not ready to draw the conclusion that general linguistic theories are not required for document retrieval. They suggest that it would be more productive to recognise both that today's linguistic theories are far from adequate and that document retrieval systems are not well understood. Linguistics should turn their attention to larger units of discourse than the sentence, and online retrieval systems are likely to involve retrievable units smaller than traditional documents.

3.1 Syntactics

The syntactic level in semiotics indicates the analysis of physical signs and the relation between them. **Syntactic analysis** identifies the role of the word in the sentence, that is, its grammatical class and the relation among the words in the sentence. The syntactic level is designed to group the words of a sentence into structural units, representing the grammatical structure of the sentence.

The use of syntactic theories in information retrieval has been explored in the following types of procedures:

the analysis of
- input documents or their surrogates
- syntactic aspects of document description
- syntactic aspects of the indexing languages

Most of these experiments have been carried out in the 1960s and they showed that syntactic information contributes little to retrieval performances. A possible explanation of the situa-

tion to date is given by Sparck Jones in her "Automatic Indexing: A State of the Art Review":

> It must be allowed that syntactic descriptions may be of value in specific contexts, but there is no hard evidence to the effect that they are generally of value ... However, it may be that the correct way of providing syntactic information in document descriptions has not yet been discovered, and that when it has, it may be worth writing, or attempting to write, programs to provide it automatically". (Sparck Jones 1974a, pp.3,9).

3.2 Semantics

In semantics the study of meaning, is implicated by the relation between signs and objects or events which they stand for. At the semantic level, the information conveyed depends on the individual's ability to perceive the relationship between the signs and the object they denote. Semantic information is independent of specific communication events.

Semantic analysis adds contextual knowledge to the purely syntactic process in order to restructure the text into units that represent the actual meaning of a text. A variety of pre-established semantic characterisations of the terms are normally used to obtain satisfactory semantic interpretations, that is, to associate words with simple concepts.

The goal of semantic analysis, as applied to document processing, is to identify the subjects and content-bearing words of the document or surrogate text. Within the field of the subject description of bibliographic items, there are three types of procedures for creating access to the documents:

a) **keyword normalisation,** that is, stripping words of prefixes and suffixes;

b) **dictionary or thesaurus reference,** in which the extracted word is looked up in a thesaurus;

c) various **classification** techniques aimed at grouping related words.

However, the value of these techniques of semantic analysis for automatic indexing is, at best, unproven; much more experimentation is needed. Although it is possible to be more definite about the value of automatic methods in relation to semantics than it is about syntax (Stevens 1970) (Doyle 1975) (Sparck Jones and Kay 1977) (Robertson 1977).

Language processing by means of semantics involves attempts to design, store and use **cognitive structures** or networks that can interrelate meanings of words in a way resembling the human brain's own ability to interrelate meanings and concepts. Though the human brain is able to store vast amounts of information relating to each word or concept, a computational linguist is likey to try to see how far he can get in simulating cognitive processes by storing a modest amount of information per word or concept. Cognitive language processing is a fascinating area, full of potential - it remains to be seen how many of the promising research ideas can be applied to an operational IR-system (Doyle 1975, p.325).

3.3 Pragmatics

At the **pragmatic level** information processes can be characterised by the interaction between the information transmitted, the interpreters or users of information, with their systems of values, and the objectives which the information serves to promote. It is the study of signs in relation to their users. It is the most general, most inclusive level of semiotics. It is related to human values, not only economical, but ethical and aesthetic values, and to human experience in a very broad sense of the word.

Human indexing has been studied predominantly at the syntactic and semantic levels of semiotics, assuming somewhat unrealistically, that the process involves only meaning for (fairly) standard interpreters under (fairly) standard conditions. The fact that the meaning of a message varies with the characteristics of human individuals who construct and interpret them, has not been fully appreciated in researches carried out in the field of machine indexing.

The indexing process implies the **understanding of the meaning of text** and the ability to **make value judgements** about its information content relative to the perceived interests of the expected user. Consequently also the mechanised indexing and abstracting processes, based on the extraction of meaning from text, should imply a deeper analysis of the information content of documents.

As it has been pointed out before (page 17-18), linguistic research contributed little to the development of appropriate analytical instruments for the conceptual analysis of document contents and for the exploration and organization of Poppers World 3, the "objective knowledge" recorded in written works.

There are critical voices against the strong syntactic orientation of language analyzers and the quantification of informa-

tion processes. They question the continued usefulness of those conventional theoretical structures on which most of today's indexing methods are built upon (Artandi 1973, 1976) (Brookes 1983).

There is a small stream of research efforts, however, which has persisted in the belief that there is an inherent, discoverable and computable connection between language structure and information. But the tools for extracting and exploiting this connection have been (for seveeral reason) slow in developing. In cases were limited tests have been so far carried out, very complex, high quality language analyzers have been used for manipulating laboratory models of the inference rules and the large universe of discourse information structures (Sager 1977) (Schank and Abelson 1977).

Researchers in the field generally agree that the empirical approaches are not sufficient here and considerable theoretical knowledge is needed before further progress can be made. It has been realized that **human natural language implies highly integrated processes and that the understanding of meaning goes beyond the linguistic context provided by the text itself.** Prominent scholars argue for that language analysis systems must necessarily be based on the global use of many different approaches at each point in the process (Langefors 1970, Section 23.17) (Langefors and Samuelson 1976, Section 3.9-10) (Sparck Jones 1979) (Schank-Lebowitz-Birnbaum 1980).

One major concern in the pragmatic study of language has been the nature of signs (words) from the point of view of communication - where the classical formulation of the "word-thought-thing" in the semiotic triad (see Fig.3.1) is interpreted not as representing the essence of meaning but rather the way signs are understood by the receiver (i.e. how he becomes informed).

Although linguists and psychologists have mainly been interested in the relation between "word" and "things", pragmatists argue that the signs are the most crucial elements in all communications - both between persons and within persons. They assert that the meaning is not considered as attribute of the data but rather as something which depends on the actual interpretation of the data by the user, therefore it seems somewhat inconsiderate to occupy oneself with meaning rather than with the way meaning is produced. Thus, here is the process of understanding rather than the product of understanding which becomes the most important phenomenon of communication (Pierce 1961) (Cherry 1966).

Another consequence of the pragmatic view is that signs do not have any meaning unless they are actually used. Consequently, signs have different meaning to different persons, or the same

data may lead to different states of information in different users - which is a very important aspect of communication.

Since indexing and the documentary information transfer imply different events of communication - based upon technical and psychological factors - the pragmatic view contributes very usefully to a more **efficient running of IR-systems.** Efficency is meant here not only as the ratio of the success in the system's performance to the costs involved, but rather in the users ability to interpret the signs in the way which was intended by the sender. It means also that the problem of effective communication becomes hereby much more extended e.g., compared with those early discussions carried out on the basis of the "classical" theories of signal transmission (Shannon and Weaver 1969).

Nowadays there are many different kinds of criteria given for an efficient communication, focusing on the user and his efforts to use the system (Lancaster 1979) (Hein 1982) (Markey 1984). It is the author's belief that arguments like these, are of great importance in our daily (computing) life.

3.4 Indexing as a communication process

Thanks to the publication in 1949 of Shannon's theory on signaling the engineering problem of information transfer has been throughly studied and is relatively well understood today. In his writings on signaling theory Shannon discussed in detail the interrelationships of Message, their Codes, and the physical limitations imposed thereupon by the physical Channel used in transmission.

The main limitation of Shannon's **"Mathematical Theory of Communication"** (1949) is that it is concerned only with physical signals and omits to take into consideration who signals to whom. In Shannon's model of a communication system the information source selects the message out of a set of possible messages. The transmitter changes the message to produce a signal that is suitable for transmission over a communication channel from the transmitter to the receiver. The receiver reconstructs the transmitted signal into a message and forwards it to the destination.

In the interpretation of meaning by the receiver as compared to the intended meaning of the sender, we have to go beyond the engineering problem of communication and take into consideration a broader range of communication problems. In Shannon's model the transmitted signal, ideally, is identical with the received signal and does not include the noise during the transmission or the unwanted distortions and errors that cause

a change in the transmitted signal. The concept of information applies here not to individual messages but to the situation as a whole, and is based on the statement that the movement of signs in space, from the transmitting system to the receiving system, follows the same physical laws as the movement of any other sort of matter-enery.

This model of a communication system became later expanded including semantic noise, as distinguished from engineering noise. E.g., it refers to the distortions of meaning or the ambiguities which are not intended by the information source but which nevertheless affect the destination (See also Miller 1968, p.406).

In spite of the great importance of the Shannon-Weaver theory (1969) for a better understanding of the process of information transfer, it is important to point out its limitations.
E.g., the theory is concerned only with the **capacity of the channel** and ignores the capacity of destination. It does not take into account, that, by direct analogy to channel overload which results in reduced fidelity, overcrowding of the capacity of the audience creates confusion and error (See also Miller 1968, p.98).

It does not take into consideration either, that when information patterns are processed over space, while being processed, are often shifted from one matter-energy state to another, from one sort of **marker** to another, and if the form of the signals does not remain relatively constant during these changes, the information is lost or distorted (Miller 1968, p.12).

To have the capability to cover **all the broad range of communication aspects,** can be considered as an intellectual challenge. Therefore, James G Miller's integrated, multidisciplinary analysis of the matter-energy and information processing functions in living systems has a great importance for studies in communication. In his analysis he tries also to generally apply to the living systems some of Shannon's and others' communication theories (Miller 1968, pp.97-98).

This approach has been useful for the author to gain a deeper understanding of how human beings receive and decode information. Through a detailed analysis of how alpha, beta and gamma-coded information is processed at the different levels of living systems, the function of Encoding and Decoding is presented in an evolutionary perspective (ibid. pp. 394-407, 438-443, 540-543, 554-555, 792-794, 824-826).

All these aspects of communication have fundamental importance also in the documentary information transfer processes where human beings receive and decode information carried by marker who move in space.

Indexing implies the understanding of the meaning of messages and the expression of this understanding in the form of new messages (index terms) that will serve as surrogates for the original messages. In this process the indexer (man or machine) functions both as a source and a designation - both as an **encoder** and **decoder** of messages.

In many existing bibliographical IR-systems, document contents are represented by index languages, in which the subject terms have assumed special meanings compared with their interpretations in natural human language form. These languages are artificial constructs, representing in a sense meta-languages in relation to the languages used in the documents. For a successful execution of a search it is necessary that both the indexer and the user follow the peculiarities of the used indexing systems, as well as that the given subject matter has been interpreted in a similar way both at the time of indexing and the search. This means that relevance judgements, the process of understanding and extracting relevant information, is based entirely on signs (words and phrases) derived from a superficial description of document content.

Although it should be noted that the application of computers to information retrieval has opened up quite new possibilities for internal and external representations in information systems. In many computerised IR-systems the representations used have simply been adopted from those used in printed tools (without an analysis of their suitability for machine-readable files). Because one form of representation is suitable for one mode of searching (e.g. manual), it is surely not adequate for another mode, therefore **existing IR-systems often are badly structured and used for the utilization of computer power.**

One of the most urgent problems in the development of present IR-systems is to detect how it is possible to enhance the searchability of computerised bibliographic files - by modifying their structure or adding extra data. Design goals for representations in modern bibliographic files have to be extended to involve the basic concepts of human communication; e.g. the meaning of messages varies with the characteristics of the human individuals who construct and interpret them (See also Miller 1968, pp.406, 97) (Hypothesis 3.3.3.2-16).

Considering the complex problems of communication and language comprehension involved in indexing, it is quite obvious that pragmatics has the greatest relevance to find the answer to the problems of effectiveness in IR-systems. (Though syntactics and semantics are not to be considered here as exclusive categories of language analysis, but to a substantial extent overlap each other and being implicit in pragmatics).

The following citat will illustrate how the **implications of semiotical problems** can serve for a better understanding of the methods of text analysis and the full communication process in information retrieval:

> "The proposition that information is always determined relative to the interpreter, his actual internal state and organisation of knowledge, allows indexing to be considered as an activity whose purpose is to provide information to an interpreter of whose ever changing purposeful state the indexer has only a minimal, generalised, and frequently static, model. Since the author wrote with the reader and not the indexer in mind, close similiarity of the indexer as an interpreter, his model of the reader, and the author's model of the reader should increase the potential for effective communication in a document retrieval system". (Artandi 1973, p.245).

4. THE COGNITIVE VIEWPOINT

The representation of information needs and the representation of the information content of texts are central issues of information retrieval. In traditional IR-systems this representation is based on various sorts of indexing and classification. Though in modern IR-systems a text is mainly represented by its title and abstract or, in few cases, by the entire text itself. The essencial representation of question lies in the practical problems involved in finding the text appropriate for answering a particular information need (Belking, et al 1982). If the text is represented by something other than an uncontrolled natural language familiar to the user, an intermediary (such as a librarian or an information scientist) is necessarily placed between the user and the system; in this situation the complexity of the representation is considerably increased. Thus, information storage and retrieval in library operations relies heavily on language and its semiotic components (See Chapter 3).

Language comprehending - which means the processing of gamma-coded, learned information in living systems, from the Organism level all the way up, - is a very complex communication process: certain patterns of input signals are transformed to other patterns that represent them in the ensemble of an internal code of the nervous system. Each person uses his own set of meanings or semantic space, even for common words, which differs somewhat from the meanings of others.

As a consequence of this, information processing is always affected by the differences in the individual's knowledge structure, which is determined by the individual and social/collective experiences, education, training, etc.Although this notion has been familiar for quite some time, its implications for our understanding of phenomena such as visual perception and language understanding have only recently been recognised.

The central point is here the **"cognitive view"** which indicates a common trend in various scientific disciplines:

> Cognitive linguistics, cognitive psychology, cognitive sociology, cognitive antropology have now a new common denominator: The black box ideology (where systems are studied simply by noting correlations between their inputs and outputs) has been replaced by the white box ideology (the study of systems as determined by the models they have constructed for the environment they are interacting with)." (Apostel 1977).

4.1 The cognitive stage in thinking on information processing

De Mey (1977) states that the adoption of the cognitive view is a recent stage in a **series of four stages through which thinking on information processing has developed.**

The parallel to IR evolution seems applicable:

- a monadic stage during which information-units are handled separately and independently of each other, as if they were single self-contained entities;

- a structural stage which views information as a more complex entity consisting of several information-units arranged in some specific way;

- a contextual stage where, in addition to the structural analysis of the information-bearing units, contextual information is required to disambiguate the meaning of the message;

- a cognitive or epistemic stage in which information is seen as supplementary to a conceptual system that represents the information processing system's knowledge of its world.

The **cognitive stage** implies a context which is produced by the conceptual information system (e.g. man) itself. The system's knowledge of its world provides the basis for decisions on which ambiguities should be eliminated.

The cognitive view is a vague and mystifying position unless it reveals how "world-views" are organised. One could characterise this position by saying: "meaning is in the world-view of the subject". It does not recognise any meaning in an isolated signal or message: the meaning is always synthesised by the receiver out of his own knowledge.

If we accept this proposition, then the meaning must be ultimately referred back to the intent of the user: beyond the formal meaning of a grammatical sentence there is the problem of associating words with the world of man's society and his organised knowledge. Thus, the meaning of language depends not only on grammatical order but also on the structure of the world around us and a range of experiences to which the words refer - which is not necessarily common to all cultures. (Pierce 1961).

Context in the **contextual stage** is still considered to be part of the surroundings of a signal, external to the perceiver or the listener and which has to be noticed together with the signal. In the cognitive view context becomes something supplied by the perceiver or listener: It is the knowledge he invokes to analyse the signal and to determine its meaning. It is the knowledge that permits him to look for and to detect very selectively those structural characteristics and those contextual

elements which yield an interpretation of the message that is congruent with that knowledge.

Applying the cognitive view to information retrieval, one can find useful explanations to the present state of IR-systems and (hopefully) also gain some hints to the further development of these systems. Thus the task of IR is to bring the cognitive structures of authors, system designers and indexers into accord with those of the information worker and the user, in order to cope with the actual need.

Collective cognitive mechanisms often described in paradigm theories also influence the structure of classification and indexing systems, and thus have implications for the relations of topics and concepts treated in the body of literary and information needs. This sociological aspect of Kuhn's paradigm concept (1970) has great relevance to the development of IR-systems because, as De Mey points out:

> "Many information science investigators have backgrounds in library science, and bibliometric analysis of scientific literature. They see themselves as also concerned with knowledge, scientific knowledge aspects, specifying paradigms in terms of their cognitive ingredients, mined from statistical analysis of its literature." (De Mey 1980, p.60).

The clustering of literature is a non-obtrusive measure of the social aspects of science. Therefore the librarians' traditional way of creating ordering systems for document collections and of providing access to the contents of documents has basically determined (and some times limited) the design of present IR-systems.

Differences in **individual knowledge structures** is one of those basic concepts which has important relevance to user-librarian communication as well as to communication in the international networks. In cognitive psychology two different groups of knowledge structures are distinguished: **semantic and episodic memory** (Tulving 1972). The first has been used to refer to the class of information characterised by the definitions that people have within their memory. The second group refers to the type of information which is about particular events that have been experienced. Semantic and episodic memory are related: both are contained in the individual cognitive structure and semantic memory usually develops from information stored in episodic memory.

The distinction between the two different ways of retrieving information from the memory has proven to be useful in discussing user-systems interactions and the different forms of search procedures undertaken by librarians (Ingwersen 1982). It is of interest to know how episodic memory (private knowledge of the

individual) influences search processes and how semantic memory (formal knowledge), shared by others and resulting from education, including training and skills, is employed in IR-systems. Our awareness of this distinction between the two classes of information in human memory can considerably improve the effectiveness of information systems - specially those which function in heterogenous user environments and on an international level. That is, to strive more in our communication, to use that type of information which is processed by semantic memory.

Davidson (1977) has examined how the users' relevance decision is affected by the fact that individuals selecting information from alternatives do it in a somewhat idiosyncratic fashion. The pattern of the empirical results confirmed the prediction that approximately 30% of the relevance-decision variance is attributable to variables tapping the breath and flexibility of the individual's cognitive structure. The **"openess to information"** has been operationally defined as a series of cognitive style variables (open-mindeness, rigidity, category width, locus of control, anxiety and defensiveness).

The value of research issues on knowledge structures and on the stimulation of cognitive style is that they begin to allow us to deal with users from a fundamentally different conceptual point of view and help to empirically recognise the **"epistemic who"** in information systems.

Luria's studies (1976) on the development of the cognitive process provides another useful contribution to a deeper understanding of classification problems in IR-systems. In the classification of objects Luria isolates two different kinds of human approaches:

i) categorial classification, i.e. a person sorts out an abstract concept and chooses the objects which can be included under this concept.

ii) situational classification, i.e. a person tries to involve the objects in different concrete situations from practical life, thereby classing objects which belong together.

This classificatory approach seems to be in harmony with Vygotsky's dichotomising of concepts into scientific and common concepts (1962) and to be similar to the frame and script concepts of Minsky (1975) and Schank (1977). When using the concept of knowledge structures, the distinction between categorical and situational classification must be implied.

4.2 Cognitive sciences

The problem of understanding how humans interact with information extends the domain of information science to include not only linguistics but also psychology, education and artificial intelligence - all confluencing about the questions of human cognitive activity. All these disciplines are interested in information and knowledge representation and natural language processing, though they use different approaches to their representation.

The questions of text (any meaningfully ordered set of symbols) and knowledge representation have recently become very prominent in these allied fields, and - on the basis of their common concern with human cognitive activity - they have been collectively called **"cognitive science".** This new field aims to study the content, structure, function and development of mental representations. It is an **integrative approach,** not merely in that it relates knowing, acting and feelings as closely interacting functions of the individual's view of the world, but also in that it has a potential for integrating all the different theoretical disciplines dealing with the mind.

Margaret A Boden, one of the mentors within cognitive science, has the following attitude to this new scientific research field:

> "Cognitive Science can counteract the insidiously dehumanising effects of the natural sciences and behaviourist psychology.... It promises to show not only how knowledge and subjectivity are possible in a basically mechanistic universe, but how purpose, freedom and moral responsibility are possible also. It offers a truly human psychology without denying our biological embodiment or losing sight of our evolutionary origins". (Boden 1977, pp.227-8).

The direct relevance of these disciplines to information science has usually been interpreted as their offering **a body of theory** which information science could make use of in solving its own problems. Many investigators in information science have suggested that information science is also concerned with the problem of cognitive science and a great number of works have studied the connection between information science and the component disciplines of cognitive sciences. (Sparck Jones and Kay 1973), (Atherton 1975), (De Mey 1980), (Harbo, et al. 1977), (Brookes 1977, 1983), (Smith 1979). Possible links between cognitive science and information science has been thoroughly explored by Belkin (1977a), who discusses to what extent contemporary approaches to the representation problem in cognitive science might be relevant to the problems of representation in information retrieval. However, there are also scientists who are sceptical of the possible contribution of the allied fields, primarily because of the lack of a sound theory which

is relevant to the problem of information science (Sparck Jones and Kay 1977), (Sparck Jones 1979). Nevertheless, there is a need for new, unconventional research approaches which are engaged in new avenues of research directly related to the human information processing problem. As Newell (1973) puts it: "sciences are not defined, they are recognised".

The emerging disciplines of cognitive sciences and knowledge engineering suggest challenging new roles for information scientists and pave the way for more ambitious goals in the development of IR-systems. **Artificial intelligence (AI)** concepts and techniques have now begun to influence the design and development of information systems. The motivation for AI applications is to shift some intelligent activity from the user to the computer, making the computer a more active partner in accomplishing the task for which the system is designed. The concept of pattern recognition, representation, problem solving and learning can provide a framework for studying AI applications in information systems. (Smith 1979).

Some designers of IR-systems have focused their attention on providing natural language access to data bases and abundant internal representations supporting inference.

This research area has good theoretical potentialities, but its results, at present, are far fetched from the practical use:

> "investigators in reference retrieval have explored a variety of techniques that can be viewed as AI applications but only have a few people explicitly referred to them as such...An awareness of and support for possible AI applications in information systems are not widespread in the information science community. It is difficult to predict how soon such applications are likely to move out of the laboratory and into large scale operational information systems". (Smith 1980, pp.94).

5. ELEMENTS OF IR-SYSTEM

5.1. The Notification hexagon

Fairthorne's classical hexagram (1967) visualizes in a network model, the interacting elements of an information system. It is not a geometric figure, rather a diagram, a conceptual model showing which elements coexist with which. These elements cannot stand on their own, but they can be studied separately in the first instance.

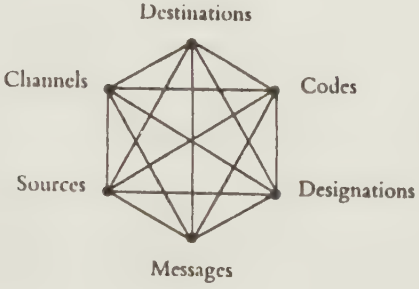


Fig. 5.1. Elements of information system.(Notification hexagon)

The hexagram contains the technical aspects of information transfer as well as the non-technological aspects of communication, giving a proper representation of the complete of communication act, considered as:

WHO = Source
communicates WHAT = Message
to WHOM = Recipient
in what WAY = Channel

As raw material library services are concerned with the activities of mentioning who recorded, in what form, where, for whom, and what he mentioned. Fairthorne's model has proved to be a useful model for analysis of the problems of library science and information retrieval, since this field is concerned with all aspects of the task of bringing the author's message to a destination, to the attention of the ultimate readers.All the physical, conceptual and logistic problems

inherent in this service task are characterized by Fairthorne as "Notification".

The **Notification hexagon** contains all the necessary elements for bringing together Message and Destination - if the direct connection between them is missing. Since in documentary information systems there is no direct route between Message and Destination we can create "flow" only by using one pair out of Source, Destination, Code and Channel to create a tetrad of which Destination and Message are the opposed pair. This is in fact what we do in practice, with author indexes, subject catalogs, shelf arrangement etc. - by creating ordering systems to the document collections of our libraries.

Messages (e.g. in documentary texts) emanate from **Sources** (authors, publishers, organisations) in Code form (i.e. in a symbol system consistent with the physical mode of communication, the **Channel**), and reach **Destinations** (readers, users), who are aided in their choice of **Message** by Designations assigned to Sources, Messages or Destinations (e.g. subject index entries, natural language terms).

The identification of about 20 various triads of the model is an instructive exercise to show which elements coexist in the different part of subsystems. These elements can exhibit "flows" in the ordinary reasonable sense if, and only if, two triads have two elements in common. E.g., if the Source and the Destination are in accord about the Code and the Channel, there can be flow from Source to Destination. If they are not, there will obviously be no flow.

Fairthorne in his "morphology" paper analysed and gave descriptive names to some of those twenty triads which are representing different subsystems within the Notification hexagon. Mooers, being stimulated by the hexagon, took up the challenge to complete the model and his paper **"Analysis of the Hexagon of Notification"** (1974) he detected and identified some more useful symmetrics that Fairthorne had not mentioned.

He has carried out several groupings of triads whereby modelling with triad pairs and their mirror-image pairs he analysed in detail certain domains of the information transfer process. In this way he could look more deeply into e.g., problems concerned with Source/Destination.

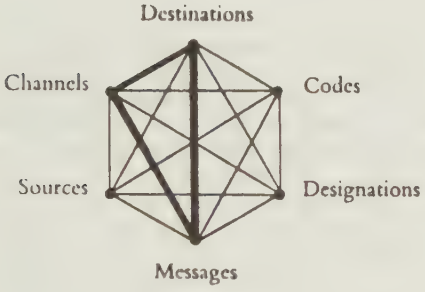
Each of Mooers' twenty triads theoretically permits six substantially different analyses. Of course, some of these 120 combinations are essentially meaningsless according to any reasonable interpretation, but in any case, the model offers an useful "check list" of combinations.

Another useful feature of the model is that by the application of Shannon's statistical analysis an "entropy" of the set of

classes can be calculated. Mooers also gives examples of partial derivative analysis where the Notification hexagon has proved to be a useful theoretical model.

5.2 The analysis of SAP-indexing by the Notification hexagon.

The author's intention here is to make an analysis of the elements of SAP-indexing based on the triads of the Notification hexagon. The analysis is limited to those triads which are particularly related to the domain of subject description of documents.



Triad A

This triad can be interpreted as dealing with "Documentation and Publication Service for Users", or as **"Document Retrieval".**

Channels are the physical means by which Messages (information) are transmitted (i.e., books, publications, wire facilities) irrespective of Coding (language) or Sources (authors).

The central idea is to identify documenthood with authenticity and reliability, and to explore the relationship of documents to people and to themselves. The time-honored way of looking at the function of a library and of bibliography, nowadays has been reexamined and the modern view shifts the burden onto bibliographic control, from keeping track of physical objects to the devising schemes for the intellectual organisation of the information resources of the library.

Concerning some basic characteristics of current document retrieval I want to point out that all IR systems now available generally have a great weakness, that is, they depend on the so called **"best-match" principle** (Belkin, et al 1982). This principle means that given a representation of a request for information (e.g. a "query" or a set of search terms"), the best possible system response will be the text whose representation matches it most closely.

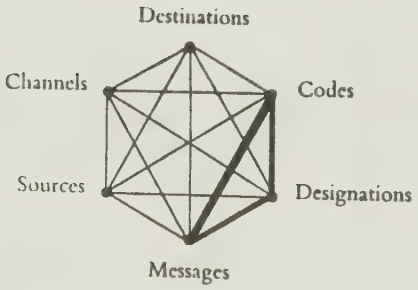
There are two assumptions underlying this principle: 1) it is possible for the user to specify precisely the information that he requires and 2) information needs (or at least expressions

of them) are functionally equivalent to document texts. Thus, the best-match principle looks first for a document which is just like the expression of need; that is, which is functionally equivalent to it.

As it is known, that in most situations the user is not able to specify all the relevant aspects of his problem, in order for the system to work optimally. Thus, the weakness of the first assumption is obvious. Consequently, the probability to find all documents treating all relevant aspects of the problem is also limited by the user's ability to specify and list these concepts.

A critical review of the second assumption bringings to light the problem of document representation in current IR systems: equivalence between expression of need and document text of its representation. That is, how well search queries respond to the representation of need as well as to the representation of the document ideal for resolving that need.

In a "second-generation" of IR system, the representation of information need is the central problem. The design of future IR systems must be based on the attempt to give an explicit structural representation of the **anomalous states of knowledge** (ASKs) underlying information needs (Belkin, et al.1982).



Triad B

In Fairthorne's terminology this triad is named **"Marking".** Here Designation acts upon Message to classify them, and Codes provide a symbolic representation of the assigned classes.
We have here an abstract system of categories, i.e. a classification system whereby elements may be grouped into subsets or classes, such as classes of elements which possess specific attributes.

Codes are interpreted here as the representation of the traditional indexing and classification routine. It is assumed that the set of symbols are predestinated and consequently followed by all users.
The triad is concerned thus with ordering systems where index terms or class codes are selected from a **"controlled vocabulary"**, i.e. from a limited set of terms that must be used

for both indexing and searching the documents in a particular system. It is the pattern of traditional retrieval tools where class relationships are expressed once and for all, and the indexer assigns to the document one or more class labels from a standard authority list of subject headings.

One of the main limitations of this kind of "Marking" in documentary information systems is that one has the ability to coordinate or correlate classes only at the time of indexing and not at the time of searching - which is to be preferred.

It is a well known fact that the user can very seldom describe precisely the information needed. In consequence, several modifications of the need formulation, normally done during a search, are based on browsing processes.

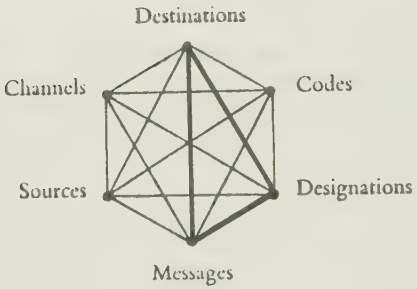
For document retrieval it is therefore preferable to use what is called "manipulative" systems for indexing. In accordance with the next Triad (C) - these systems provide freedom and flexibility to manipulate classes at the time of search. It means that the searcher does not have to be guided in the selection of search terms by a predetermined and rigid system of categories.

SAP-indexing represents a new way to produce subject descriptions of monographic publications. It is an approach to eliminate the known shortcomings of existing standard library classification systems, by which the varied and extensive contents of a book is described by means of limited numbers of classification codes.

SAP-indexing method is based on the philosophy of an extended use of human natural language in IR-systems. It means that there is no vocabulary control either in indexing nor in searching; subject description is made by an unlimited set of terms drawn from natural language - i.e. the author's own vocabulary. This type of **"free-text"** search technique assisted by a powerful retrieval system, facilitates many exploratory strategies in online searching.

Striving after refined methods for document representation, the Codes in future IR-systems may be quite different from the present thesaurus and classification systems.

Natural language input compiled with e.g., the "search only" loosely structured thesaurus, or creating some type of master switching language (that will make a number of different vocabularies compatible with one another), may become the conventional approach to "Marking" in the future (Lancaster 1980).



Triad C

This triad is named **"Prescription"** and it means the act by which a prospective user of information describes the information needed. In that sense, Designation operates upon Messages to set them into classes, for the purpose of the Destination.

Mooers presented another "obverse" interpretation in which the Destinations are put into classes by Designations for the purpose of Messages. This second interpretation would seem to be operative when an automatic distribution system is set up and this categorizes the recipients' needs.

Here the Destination-to-Designation link seems to be paramount. In the present author's interpretation the triad is concerned with the problem of user orientation in the development of IR-systems. This triad indicates new client-centered roles for librarians which would phase out those traditional roles, where most energy is directed towards the handling of graphic records as mainly physical objects. (See also Chapt.2). It implies that the social purpose of librarianship, are shaped by the nature and character of the group served (Shera 1980).

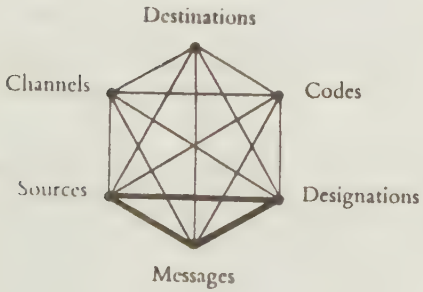
It would also change the view and the demand on the bibliographic control mechanisms: because of the prevailing use of computer technology, bibliographic information is becoming more and more directly accessible to almost anybody who needs it. This fact is placing a new demand on retrieval tools: they must be formalized so that they facilitate search and identification for non-bibliographers too.

Information systems must offer bibliographic control at different levels, i.e. physical and intellectual access to recorded knowledge for different needs of information. Today indexes, catalogs and data bases still, to a great extent belong only to the information specialists' **"epistemic community".** (The term is Burkart Hotzner's (1968) and it means the social context in which our knowing and our ways of knowing have meaning and validity).

Contrary to other methods of deep indexing, SAP-technique does not require any specific subject knowledge on the part of the

indexer: for indexing, the terminology given by the author is used, and the index terms are selected from the table of contents, the back of the book indexes. The criteria for selection are easy to understand and easy to apply and SAP abandons a great many of the traditional rules for professional indexing. The user is allowed to form the query in natural language, which means that it is possible to use very specific terms, and also combine them freely into phrases during the search process.

The qualified knowledge and the experiences of an indexer are partly replaced here by highly developed automated routines where searches are executed in man-machine dialogs. Considering the trends in costs of library services, SAP is an example of how nontraditional thinking and low-priced technology can help in the development of a new service that was not previously economically feasible.



Triad D

Similar to the previous Triad (C), Designations operate upon Messages without having a systematic notation implied for the classes so formed. The third element of the triad is Sources, which implies that the categorizations are formed from the viewpoint of the Sources or the authors. Therefore this triad can be called **"Subject Classification from Authors' Viewpoint".**

In contrast to Triad (B), where the classification of Message is provided by a prescribed symbolic representation of the assigned classes (the codes) here it is guided by the person or physical agent who has produced the "writing". The titles, abstracts, key words, etc., which are normally used for document representation in data bases containing periodical literature, can also be looked at as a certain form of subject classification given by the author.

Here we can find corresponding elements to the basic idea of SAP: for indexing and subject description of books use the author's terminology and expressions found in the book! Instead of using some kind of meta-language for classification and indexing, it seems to be a more natural and obvious way to make use of the "inherent" attributes of the book for retrieval.

Since experience shows that it is very difficult to find and define the "best" subheadings or classification codes for a given document, it seems to be preferable to produce subject descriptions of books taken from the books themselves (Wormell 1984a Section 1.3, 1.5).

By selecting terms and phrases from the tables of contents, the back of the book indexes, and also from the nontextual material such as tables, graphs, photographs, charts etc., we augment considerably the traditional subject description of books.

SAP-indexing does not only mean the possibility to produce detailed subject description of books at a moderate cost, but also from the **cognitive viewpoint** a more precise matching between the information needed and the content of the retrieved document. Using the author's words and terminology for document indexing, the Message is processed and mediated in a **"direct"** way to the user, compared with the traditional routines for indexing and classification where the subject content of the document is coded and recoded into a meta-language via the indexer's intellectual analysis.

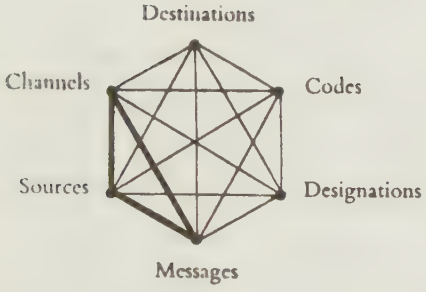
The cognitive view states that any processing of information, whether perceptual or symbolic, is mediated by a system of categories or concepts which for the information-processing device, are a model of his world - determined by his individual and social collective experiences, education, training, etc. (See also Section 4.1).

Considering the importance of the **individual knowledge structures** in information processing, the role of intermediaries in IR-systems potentially implies a barrier which can distort the "flow" of information between the Source and Destination (Wormell 1984b).

The relatively simple and evident way to subject description which is implied in SAP-indexing, opens new perspectives in the development of future online library catalogs. Used in the combination of some kind of ordering system, it represents a really augmented and multi leveled subject analysis of document contents. By adding SAP-index terms to traditional catalog entries, we produce detailed subject description which, at a moderate cost, makes it possible to tap the rich resources of our libraries' book collection - in a way which was not feasible with the traditional retrieval tools.

This indexing method permits one to search, not only a broad subject area, but also the specific aspects of a book's treatment of that subject - and that is a particularly important feature of indexing monographs!

Hereby we have the point in common with the next Triad (E) where the indexing problem of monographs is implied.



Triad E

This triad is interpreted as being concerned with the **choice of Channel used by a Source in the circulation of particular Messages.** E.g., what journal to publish a given paper in or the decision whether to write a journal article or a book about a given subject.

It is aimed to point out here the traditional differences in information contents of journal articles and monographic publications: **monographs** provide a new synthesis of previously published materials, whereas serials usually contain new materials.

Nearly all monographs are complete in themselves, whereas serials are on-going works.

Looking for an efficient way to describe the contents of the books, it is necessary to realize the fact that **indexing methods have to accomodate the content and the form of the specific work being indexed** (Wormell 1984a. Section 1.3). Therefore indexing features for monographs must differ from those which are typical for serials: the indexer of journal articles looks principally for new subjects, the monographs indexer seeks emphasized subjects.

While there are many computer-based abstracting and indexing services that provide detailed subject access to journal articles, little has been done to improve retrieval capabilities for the user of monographs.

SAP-indexing is particularly concerned with the subject description of the contents of the books. It is an approach to offer a solution to one of the crucial problems of library catalogs, the provision of a multiple access point for a multi-aspect document.

SAP has been developed in the spirit of looking for new ideas for subject description of **monographs.** Because of a lack of efficient retrieval tools in our libraries, many of the

authors' messages will never be brought to the attention of the readers. A major part of the rich contents of a book remains "hidden" on the shelves in the library.

The demand for refined retrieval tools in today's libraries is obvious. For an effective use of library resources we have to find improved retrieval capabilities even for those specific sections of the books which in traditional IR-systems are not normally accessible, such as tables, graphs, photographs, charts etc.

6. EXPERIENCES AND FINDINGS OF WORK WITH SAP

Most of the experiences and findings related to SAP are described in detail in the three comprehensive reports published at Lund University Library (Wormell 1979,1981,1983). Here will be mentioned only some of the results, which in my opinion could be of interest for some other catalog developer in Sweden.

6.1 Following a chronological order, I first want to point out those difficulties which have been connected with the **Swedish translation and application of the SAP-indexing rules** (Wormell 1979). MARC-like catalog records in LIBRIS do not have either classification codes or the Library of Congress Subject Headings (LCSH). This fact was of importance in the Swedish application of SAP and placed some limitations on the design of the two SAP-files (e.g., the different local classification codes).

Since the vocabulary in the field is changing rapidly in connection with online technology, it was difficult some times to find the proper Swedish interpretation of the SAP-specific concepts. For easy reference in the production of the indexing manual, I have used colored pages for the definitions and have also given Swedish examples for them.

Rules for indexing the captions of tables and diagrams originally were not included in the SAP-indexing manual. Therefore I had to produce an additional guide for them in connection with the design of SOU-data base (Wormell 1983. App.A).

6.2 The two experimental SAP-files have been implemented in the search systems IMDOC, 3RIP, NewIMDOC. The first two were entirely different, which means that the effects of SAP-indexing could be tested against the capacity of at least two different search systems (Wormell 1980, pp.21-37) (Wormell 1983, pp.33-49). This is of considerable value to the project, and draws attention to the fact that **the efficient use of SAP very much depends on the free-text search facilities of the given system.**

6.3 Collaborating with the programmers, at every stage of the design work, I had to find out the "optimal" solutions for an efficient use of SAP in all three systems. It was a demanding but very profitable task, which gave me a good opportunity to find out the critical points, weaknesses and strengths of computers and systems.

To exemplify how SAP functioned in the different systems I want to refer to the **"focusing"** feature such as FORMAT 3 in

the output from 3RIP and to the use of the FINN and PROX commands in NewIMDOC (Wormell 1982,pp.6-8)(Wormell 1983, pp.38, 45-46).

6.4 To avoid double work in cataloging at the creation of the SAP-file I decided to use in LIBRIS already existing catalog data. This solution caused some practical problems in the choice and identification of items in the SAP-file and was also an important limitation of the project: all items indexed with SAP must have a corresponding catalog record in LIBRIS.

But it was **an interesting experiment** because it was the first time that catalog data from LIBRIS have been tapped and converted into another online catalog outside the system. Using a special program, the most essential bibliographic information was copied out from the Swedish MARC-like catalog records (SWEMARC), and via the ISBN numbers coupled together with the corresponding SAP records (Wormell 1980,pp.17-20).

6.5 At the end of the 1970s **PRECIS** had a central role in the development of computer-aided bibliographies. (PREserved Context Index Systems represents a syntactic approach to computer-assisted indexing). At that time many Scandinavian librarians considered PRECIS as the possible solution to our catalog problems (Hedberg 1976). Naturally it was a challenge for me to **compare PRECIS and SAP.**

In 1977 the Library at Chalmers Technical College in Gothenburg, started a project for indexing all literature by PRECIS at the Center for Environmental Studies. I took the initiative for collaboration between the projects, aimed at carrying out evaluation studies comparing the two different indexing techniques.

This aspiration added a lot of extra work to my project in Lund and was one more limitation of the creation of SAP-file: all items indexed by SAP had to be identical with the material existing at the Center for Environmental Studies, as well as with the LIBRIS catalog records.

Faced with many difficulties I have finally succeeded in building up the SAP-file according to the collaboration plans with Chalmers. The 400 English and Swedish monographs have been carefully selected and were considered as a basis for evaluation.

However, the objective of PRECIS at Chalmers was never completed and in 1980 the project came to an end, having registered only <u>some</u> parts of the material

(Energiregister). Thus, the planned evaluation studies in controlled mode, with parallel searches in PRECIS and SAP could never take place and a large part of my efforts were unfortunately wasted.

From 1979-80 I carried out a study on my own to compare the efficiency of SAP-file with that of the (rather small) printed PRECIS index which was produced at Chalmers. The two indexes contained nearly the same items, monographic publications within the field of environmental hazards and energy production.

The study is presented in the interim report of SAP (Wormell 1980, pp. 38-46). Without having access to controlled test results, this comparision shows some features of the two different indexing methods, according to values such as, indexing consistency, exhaustivity, precision and currency.

6.6 The SAP-file has been demonstrated several times for the most varied public and these occasions brought useful feed-back to my work. One of these useful contacts, outside the library world, were the researchers at the Environmental Studies Program, Lund University. They helped me with query formulations as well as with the evaluation of search results. Some of them were present at some searches and gave advice and comments on the capabilities of SAP.

I also discussed with them the idea of using some kind of aids or thesaurus for the formulation of effective search strategies, because it has been discovered that the ambiguity of words and phrases, selected directly from the human natural language, places a great intellectual burden on the part of the searcher.

Discussions with these researchers verified the theoretical basis of SAP-indexing, that is, when using the authors' own vocabulary for indexing, one can provide subject description of books in a form which represents the language of discourse and the professional jargon of a particular field. Consequently, **in searching for information on specific topics, it is preferable to use the language in which the practitioners of the field write and communicate, instead of other meta-languages.**

6.7 At the beginning of the project, Autumn 1979, my tutor Björn Tell and I tried to introduce the ideas of SAP for **the staff of the Catalog Department at Lund University Library** and asked for voluntary participation in the development of the SAP-indexing technique in Sweden. There

were some catalogers who, with interest and enthusiasm, started to learn the new technique and under my supervision practised it on some books (2-3 hours a week during a short introduction period).

To give instructions and coordinate the different activities around this collaboration with the staff members meant a great burden of extra work on my part. However, I felt that it was desirable to get feed-back from colleagues and therefore I worked very hard to keep up the interest in those who wanted to practise SAP-indexing.

Checking all the indexing, done by the others, through the analysis of mistakes, I was able to discover the difficult points and shortcomings of the rules. It was a very interesting experience to see **how difficult it was for the professional catalogers and indexers to use the human natural language for subject description;** they were very much accustomed to the use of the rigid and static classification schemes in labeling subject content. Some of them felt that SAP-indexing was not a "real" professional method and they did not like the fact that the "loosely formulated rules" allowed so much individual judgement in the selection of index terms. They made the easy and straightforward rules complicated and they had difficulty in discovering the advantages of SAP.

In connection with this experiment, it is important to comment upon the fact that later on when I had to train clerical staff in learning SAP-indexing procedures, everything always went smoothly and quickly. For non-professionals, it seems to be a simple and "natural" way to select the content-bearing words and phrases on the basis of quantitative rules (such as, choosing entries if they cover 5 or more pages) and on the semantic context of the chapter headings and subheadings. After a short period of training they were capable of carrying out SAP-indexing alone and under my supervision, input it to the system.

Sometimes there were difficult items, when even for me it was hard to find a satisfactory solution, but generally it can be stated that **SAP-indexing can be carried out by clerical personel** and it has been shown that it is an inexpensive method of providing detailed subject access in online catalogs.

6.8 It was found and also measured, that **the selection procedure** for one book takes about 8-9 minutes calculated on the indexing of 35 books (Wormell 1979, APP.F). The typing (input procedure) has been estimated for almost the same time.

In the SOU-file, where the records were considerably larger (about 4000 characters per item), the average time for the selection procedure was 16 minutes and for the typing 17 minutes (not including all the administrative and practical details in the delivery of material, copying, registering items, tapes etc.).

On the basis of 1000 reports, indexed by SAP from the SOU--series, it was calculated that one record in the SOU-file generally consisted of

28 caption for tables
14 captions for diagrams
14 chapter headings (selected from the table of contents).

Cost and benefit effects of SAP-indexing are analysed in detail in my reports (Wormell 1981, 1983) as well as in the final report of the SAP-project at Syracuse (Atherton 1978, pp. 16-17).

6.9 Since one of the basic question in the application of SAP-indexing was, "What is the best or adequate size of a subject description to augment the traditional library catalog entry?". Colleagues at Syracuse have computed a **quota formula** for selection of subject description. It has been worked out and given with many reservations **to use it arbitrarily** as a rule of thumb: it recommended an average of 30-35 selections (subject description entries per title). The size of the book and the quality of the content page and index entries would force the selector to go above or below this average, but rarely double or half it.

The loose application of this quota, resulted in an average of 1200 characters being added to the traditional library entry. It has created on an average 500 access words (not unique) for each item in the BOOKS data base at Syracuse (Atherton 1978, pp. 13-16).

My impression is that the quota system is difficult to apply and probably inadequate today. I recomend therefore that **the length of description and the rules for selection need to be reviewed critically at each new application of SAP.**

6.10 Another critical point in the application of SAP-indexing (and generally the use of free-text indexing technique in library catalogs), is the question of **terminological consistency** in certain disciplines.

The most sceptical voices maintain that the terminological consistency, that is, the stability of relationship between a concept and corresponding term, is so weak in the "softer" sciences that it hinders a sensible use of all kind of indexing and retrieving in free-text mode.

In my opinion, **the combination of free-text and controlled vocabulary searches is generally an efficient way to carry out subject searches** with a satisfactory result, even with those disciplines which are mainly dealing with abstract concepts.

To help to **interpret the question of vocabulary consistency with more nuances** that is usual, I want to refer to the results of Bonzi's investigation: her study supports to some extent the hypothesis that the terminology of concrete disciplines is more consistent than that of abstract disciplines. Although the concretness or abstractness of a discipline is maybe only a contributory factor in terminological consistency and other factors, e.g., rigor of methodology, probably exert greater influence (Bonzi 1983).

6.11 As has been mentioned before, the selection of items indexed in the SAP-file has been guided by the existing literature at the Center for Environmental Studies at Chalmers in Gothenburg. Naturally, most of the books were in Swedish and English but some other languages have been represented too.

Since PRECIS has been developed on the basis of the syntactic conditions of English, the subject description for all books indexed by this technique, must be given in English adapted to PRECIS vocabulary.

In the designing of SAP-file in Lund, the **multilingual** aspect of the literature has been solved in the following way: the data base contained only Swedish and English items and all search queries had to be formualated in either of these two or in both languages.

Index terms for Danish and Norwegian books have been translated or translettered into Swedish which has been chosen as the common form for input and output in all Scandinavian literature. This turned out to be a satisfactory solution, thus the English and Swedish words have quite different spellings and the chances of getting them mixed up are small in computerized systems.

When discussing the multilingual aspects of SAP with linguists, it has been predicted that the inclusion of a third language in the data base, e.g., German, could cause

confusion and "noises" in the output, regarding the fact, that the spelling of Swedish and German words is quite similar and both languages have a considerably high percentage of compounds in their vocabularies.

6.12 **Linguistic characteristics have great importance in the use of automated search systems** and they must be carefully examined both in the design and the use of modern retrieval tools. For effective use of free-text search techniques it is necessary to have some knowledge of basic linguistic conditions; the use of truncation (left and right), positional operators, establishment of stop list, etc. are some of those areas where special attention should be paid to the language used in the documents.

Since many of the intermediaries work exclusively with the English language when using international data bases, I have noticed a lack of consciousness about the changed linguistic conditions when they are transferred to work with other languages. To draw attention to this problem in the development of effective research strategies, I devoted a whole chapter to this subject in the report of SOU-data base (Wormell 1983,pp.20-33).

In connection with the use of SOU-data base I have observed that the **stilistic characteristics of the official Swedish language** (Kommittesvenskan, SOUsvenskan, Kanslisvenskan) very much influenced the effective use of the search system. Phenomenons, such as, the exaggerated use of nouns (substantivsjukan) or the extremely long compounds (e.g., byggnadskostnadsstegring, köpkraftsansvällning) are language attributes which have to be taken into consideration when searching for specific information written in this genre. When presenting the main linguistic characteristics of this style, I pointed out their effects on modern IR-systems and gave suggestions on how to deal with them in the formulation of search queries.

6.13 In view of present catalog development in Sweden, where we are just beginning to implement a modest online environment in our libraries, there is a great demand to find (for Swedish conditions) **suitable forms for improved subject access to bibliographic records.** Therefore the results of the project should be looked at against the background of this situation.

Through the development of SAP-indexing technique I have gained insight into the **micro and macro functions of bibliographic work** and the shortcomings of traditional indexing and subject cataloging. To contribute towards

a deeper understanding of the nature and purpose of bibliographic work among librarians, I have interpreted the question of **accessibility as a tripartite conception** (See Chapter 2).

I am convinced that in searching for efficient ways to enhance subject access in our catalogs, librarians ought to break away from their intellectual conventionalism, and by an integrated approach assemble facts from relevant research there e.g., the complex issues of thought and communication are discussed.

Concerning today's controversial question about the interface between **librarianship** and **information science,** in my opinion, librarianship is a service profession and its internal service variants are shaped by the nature and character of the group served. To fulfill this task, however, it is necesarry to understand man in relation to himself and his environment as well as to have insight into properties, behavior and flow of information. Thus, I agree with Shera (1968) who states that "librarianship may not have a theoretical base yet, but it is striving to derive one from its contributory disciplines, and its is the theoretical base for the practice of librarianship".

It should be noted that until now, most of the **research work on subject retrieval has not been done by librarians,** but by specialists in computer science, linguistics and most of them have published their results outside traditional library literature.

Writing on subject access has been dominated by the system analyst's view, focusing on indexing and cataloging as primary system operations. Searching was understood as something the system does for users, and the **complex issues of user-system interaction,** with or without intermediary, were often ignored in these discussions.

My theoretical work carried out within the range of SAP represents another approach: the role of bibliographies and catalogs is seen as a key instrument in knowledge communication - where the process of human thinking and mental state is the heart of the matter.

7. SUGGESTIONS FOR FURTHER WORK

According to Miller's views on the objectives of scholarly strivings (See the citation in Chapter 1); the author also feels that the results of SAP-S are only "a fragment of af a larger pattern", which in this case is to make graphic records as useful to society as far as is humanly possible.

This purpose of librarianship implies the use of retrieval tools in serving various groups of people in different environments. In this context, SAP is an approach to tap the rich subject content of books in existing online library catalogs.

SAP-indexing has also been succesfully applied outside Lund University Library and it turned out to be a useful "hint" for catalog developers in many different environments. It has also passed certain modifications when it was used as a model for building registers/catalogs of internal and external information in some libraries within industry and public administration.

Since this indexing technique does not require the qualified knowledge and experience of a professional indexer, I am convinced that **it is a particularly useful approach to building subject catalogs in developing countries.** It is also the opinion of other librarians working in different parts of the world with whom I have been corresponding and discussing some interesting ideas on the application of SAP in third world environments. As regards the multiplicity of cultures, languages and indexable terms, these places would be ideal laboratories to see how SAP-indexing could be used to enrich conventional subject catalogs outside Anglo-Saxon library traditions.

The use of OCR technique in connection with the SAP-indexing is another area for further developments in catalog building (Wormell 1984a, Section 2.2).

Through working with the SOU-series, the author discovered that it would be possible to develop completely automatic input routines for indexing of this kind of well structured, uniformed monographic publications which are represented by the SOU-reports. Special circumstances, the table of contents as well as the captions of tables and graphs have standardized forms and their publisher (Liber) is utilizing computerized type-setting-equipment in their production, making it possible to tap directly indexing data from the machine-readable production tapes into the search system.

These ideas have been presented for representatives of both DAFA and Liber, and as far as the author knows, they are

planning on a collaborative basis, to test the technical and practical conditions for the realization of this suggestion.

This idea is opening up new perspectives in the creation of data bases where the trend is to put more and more textual information into IR-systems. Having the full text in searchable form is already a reality in certain types of IR-systems, e.g., law information and legislative data.

However, full text representation is not a sensible solution in most existing library catalogs, I want to emphasize that this suggestion of **a direct input from the type-setting-tapes into a search system, can considerably improve search effciency** also in the more conventional online catalogs. To have the possibility to display the Table of Contents, Introduction, Summary, Abstracts, Index and some other parts of retrieved documents must be regarded as a great **help in the user's relevance judgement** when interacting with the IR-system.

While progress is very slow in making the library world aware of the potentialities of this suggestion, strangely enough, a commercial company in the USA, Superindex Inc. in 1983 started a profitable information service based on it. The **SUPERINDEX data base** is available via BRS and it is providing specific subject access to the books produced by the (33) biggest international book publichers in the field of Science, Engineering and Medicine. The data base entirely consists of the back-of-the-book-indexes converted from the publishers' tapes.

Superindex Inc. has currently exciting information programs under consideration and development, such as, Superindex-journals, Superinfo, Superhandbook, Superatlas - all to build up a high level reference service for professionals.

As regards the development of all the new and sophisticated retrieval tools today, it is strange that **reference work in libraries is still believed to be basically a human process.** It is true that most librarians trained in this field are very skillful information seekers,but today the rapidly emerging requirements for more subtle and detailed access to information, require the ability to also use unconventional retrieval tools. Therefore the librarians' challenge today, facing these new conditions, is just to welcome the opportunity to become more professional in their work.

Thus free-text searching online is here to stay and it is giving the impetus to examine and challenge our premises for bibliographic work and reference services.

However, it is important to point out that only technological changes are not enough for the improvement of IR-systems. Classification, indexing and retrieval procedures should be revised and based on a better understanding of human communica-

tion than is experienced today. The information-seeking behavior of the user is another problem area which ought to be explored more thoroughly.

(The state of research in this field is rather poor in Sweden.) It is desirable that further discussions on catalog design and improved retrieval services in Sweden will focus on these needs and arouse the librarians' professional interest in exploring these potential areas of investigation.

LIST OF REFERENCES

Aspostel, L.: "The cognitive point of view: Introduction to a discussion". M. De Mey et al (eds.), CC77 - International Workshop on the Cognitive Viewpoint, Ghent, Belgium, University of Ghent, 1977, pp. IX-XV.

Atherton, P.: "Research in information science: An assesment." A. Debons and W. J. Cameron (eds.), Nordhoff, Leyden 1975, 665-683.

Atherton, P.: "Improved subject access to books in online library catalogs." Proceedings of the 1st International Online Information Meeting. London, 1977, 131-138.

Atherton, P.: "Books are for use. Final report of the Subject Access Project." Syracuse University, School of Information Studies, Syracuse, N.Y., 1978.

Atherton-Cochrane, P.: "A Paradigm shift in Library Science". Information Technology and Libraries, 2(1983)Dec, 3-4.

Atherton-Cochrane, P.: "Modern subject access in the online age." (Continuing education course). American Libraries, 14(1984)Febr. - May.

Artandi, S.: "Book Indexing by Computer." Ph. D. dissertation. Graduate School of Library Service, Rutgers State University, New Brunswick, New Jersey, 1963.

Artandi, S.: "Information Concepts and their Utility." Journal of ASIS 24(1973)4, 242-245.

Artandi, S.: "Machine Indexing: Linguistic and Semiotic Implications." Journal of ASIS, 27(1976) July-Aug. 235-239.

Axelsson, G. et al: "Förstudie avseende ämnessökning i information fran Libris och Bums databaser med IR-system." AU-gruppen AB: Stockholm, 1981.

Austin, D.: "The development of PRECIS: a Theoretical and Technical History." Journal of Documentation, 30(1974)1, 47-102.

Bakewell, K.: "The value and use of book indexes." Library and Information Research News (LIRG), London, 2(1979)7, 10-11 (1979a).

Bakewell, K.: "Indexers' reaction to PRECIS." Journal of Documentation, 35(1979)3, 164-178.

Bates, M.: "Rigorous Systematic Bibliography." RQ XVI, 10.

Bates, M.: "Factors affecting subject catalog search success." Journal of ASIS, 28(1977)3, 161-169.

Belkin, N. J. and Robertson, S. E.: "Information science and the phenomenon of information." Journal of ASIS, 27(1976), 198-204.

Belkin, N. J.: "Linguistic and Cognitive Models of Information and State of Knowledge." BLRDD Report No. 5381. London, British Library, 1977, (1977a).

Belkin, N. J.: "A concept of information for information science." Ph. D. Thesis, University of London, 1977. (1977b)

Belkin, N. J. and Robinson, S. E.: "Authors' reply to Letters to the Editor." Journal of the ASIS, 29(1978), p. 212.

Belkin, N. J., Oddy, R. N. and Brooks, H. M.: "ASK for Information Retrieval." Part I: Background and Theory, Journal of Documentation, 38(1982)2, 61-71.
Part II: Results of a Design Study, Journal of Documentation, 38(1982)3, 145-164.

Besant, L.: "Online catalog users want sophisticated subject access." American Libraries, 12(1982)March, p. 160.

Bett, C. E.: "The Subject Access Project: A comparison with PRECIS." The Indexer, 11(1979)3, 145-148.

Boden, M. A.: "Cognitive Science: an integrative approach to the mind." M. De Mey (ed.), CC77-International Workshop on the Cognitive Viewpoint. Ghent, Belgium, University of Ghent, 1977, 225-228.

Boden, M. A.: "Minds and Mechanisms: Philosophical Psychology and Computational Models." Harvestor Press/Cornell University Press, 1981.

Bonzi, S.: "Terminological Consistency in Abstract and Concrete Disciplines". Ph. D. dissertation. University of Illinois at Urbana, Champ. 1983.

Borko, H. (ed.): "Automated Language Processing." John Wiley & Sons, New York, 1967.

Borko, H. and Bernier, Ch. L.: "Indexing Concepts and Methods." Academic Press, New York, 1978.

Brodda, B.: "Document retrieval - a "topological" problem." Statistical Methods in Linguistics, (1970)6, 3-13.

Brookes, B. C.: "The developing cognitive viewpoint in information science." M. De Mey et al (eds.), CC77 - International Workshop on the Cognitive Viewpoint. Ghent, Belgium, University of Ghent, 1977, 195-203.

Brookes, B.C.: "The foundations of information science." Part I. Philiosophical aspects, Journal of Information Science, 2(1980)125-133.
Part IV. Information Science: the changing paradigm, Journal of Information Science, 3(1981)3-12.

Brookes, B. C.: "Research in information science: a progress report." Journal of Information Science, 6(1983), 51-57.

Carlson, W. H.: "The Holy Grail Evades the Search." American Documentation, 14(1963), 207-212.

Cherry, C.: "On Human Communication." 2nd ed. M.I.T. Press, Cambridge, 1966.

Cutter, Ch. A.: "Rules for a Dictionary Catalog". 4th ed., rewritten, Govt. Print. Off., Washington D.C., 1904.

Dahlberg, I.: "Multi-level semantic and formal subject representation of document contents." Theoretical basis of information. Collection of papers, FID/RI Study Committee, Moskva 1974, VINITI, 1975, 323-334.

Dahlberg, I.: "New trends in classification." Peter J. Taylor (ed.), New trends in documentation and information. Proceedings of the 39th FID Congress, Edinburg, 1978. FID Publication 566, ASLIB, London, 1980.

Davidson, D.: "The effect of Individual Differences of Cognitive Style on Judgements of Document Relevance." Journal of the ASIS, 28(1977), 273-283.

De Mey, M.: "The cognitive viewpoint: its development and its scope." M. De Mey, et al (eds.), CC77 - International Workshop on the Cognitive Viewpoint, Ghent, Belgium, University of Ghent, 1977, pp. XVI-XXXII.

De Mey, M.: "The relevance of the cognitive paradigm for information science." O. Harbo and L. Kajberg (eds.), Theory and application of information research: Proceedings of the Second International Research Forum on Information Science, Copenhagen, 1977. Mansell, London, 1980, 48-61.

Doyle, L. B.: "Information Retrieval and Processing." Wiley-Melville-Becker & Hayes, New York/Los Angeles, 1975.

"Emnedata i online-alderen." Danmarks Forskningsbiblioteks-förenings Internatmøde, Nyborg Strand, februar 1984. Bibliotekscentralen, Köpenhamn, 1984.

Enser, P.G.B.: "An investigation into the automatic classification of book material represented by the back-of-the-book indexes." Journal of Informatics, 2(1978)1, 17-23.

Enser, P.G.B.: "An investigation into the automatic classification of book material represented by index list." Journal of Informatics, 3(1979)1, 37-39.

Fairthorne, R.A.: "Morphoplogy of Information Flow." Journal of the Assoc. Comp. Machinery, 14(1967)4, 710-719.

Fairthorne, R.A.: "Content analysis, specification and control." Annnual Review of Information Science and Technology, 4(1969), 73-109.

Fairthorne R.A.: "Information: one label, several bottles." A.Debons and W.J.Cameron (eds.), Perspectives in Information Science, Nordhoff, Leyden, 1975, 65-73.

Farradane, J.: "Concept Organization for Information Retrieval." Information Storage and Retrieval, (1967),3, 297-314.

Garfield, E.: "Information science and the information-conscious society." Journal of ASIS, 22(1971), 71-73.

Garfield, E.: "How will the new technology change the characteristics of libraries and their users?" S.Schwarz and U.Willers (eds.), Knowledge and Development. Festschrift for Bjorn Tell. The Royal Techninal Institute Libraries, Stockholm, 1978.

Garfield, E.: "Citation Indexing. Its theory and application in science, technology and humanities." Orig. published in 1979 by John Wiley & Sons Inc.,Reprinted in 1983 by ISI Press.

Gratch, B., Settel, B. and Atherton, P.: "Characteristics of book indexes for subject retrieval in the humanities and social sciences." Syracuse University, School of Information Studies, Syracuse, 1977.

Greg, W.: "Bibliography - A Retrospect." The Bibliographical Society, 1892-1942: Studies in Retrospect. London, 1945.

Halldén, S.: "Nyfikenhetens redskap". En bok om kritiskt tänkande om inom vetenskapen och utanför. Studentlitteratur, Lund, 1980.

Harbo, O., Ingwersen, P., Timmermann, P.: "Cognitive processes in information storage and retrieval." M. De Mey, et al (eds.), CC77 - International Workshop on the Cognitive Viewpoint, Ghent, Belgium, University of Ghent, 1977, 214-218.

Hartley, Dick: "Improving subject access to books." New Library World (1983)1, 21-23.

Hedberg, S.: "PRECIS nagot svar för oss?" Bibliotekssamfundet meddelar, (1976)1, 3-5.

Heine, U.: "Natural and Artificial Communication - Some Reflextions." LITH-MAT-R-82-10, Software Systems Research Center, Linköping Institute of Technology, Linköping, 1982.

Hildreth, Ch.R.: "Online Public Access Catalogs: The User Interface." OCLC, Dublin, OH, 1983.

Hollnagel, E.: "Some psychological requirements to a theory of information science." A. Debons and A. G. Larson (eds.) Information Science in Action: System Design. Proceedings of the NATO Advanced Study Institute of Information Science, Crete, 1978. Martinus Nijhoff Publishers, The Hague, 1983. 172-185.

Horton, F.: "How to Harness Information Resources: A Systems Approach." Association for Systems Management, Cleveland, Ohio, 1974.

Hutchins, W.J.: "The concept of 'aboutness' in subject indexing." Aslib Proceedings, 30(1978)5, 172-81.

Ingwersen, P.: "Search procedures in the library - analysed from a cognitive point of view." Journal of Documentation 38(1982)3, 165-191.

Jones, K.P.: "How do we index?: A Report of some ASLIB Informatics Group activity." Journal of Documentation, 39(1983)1, 1-23.

Kelly, Grace: "Classification of Books: An Inquiry Into its Usefulness to the Reader." New York: H. W. Wilson, 1937.

Kessler, M.M: "Bibliographic Coupling Between Scientific Papers." American Documentation 14(1963), 10-25.

Klingbiel, P. H.: "Machine-aided indexing of technical literature." Information Storage and Retrieval, (1973)9, 79-84.

Kochen, M.: "Integrative Mechanism in Literature Growth." Preface. Greenwood Press, Westport, London, 1974. (1974a)

Kochen, M.: "Principles of Information Retrieval." Melville Publishing Company, Los Angeles, 1974. (1974b)

Kuhn, T.S.: "The structure of scientific revolutions." 2nd ed. University of Chicago Press, Chicago, 1970.

Lancester, F.W.: "Information Science. Perspective Paper in Workshop on Linguistics and Information Science." D. E. Walker, et all (eds.) Natural Language in Information Science. FID Publication 551, Skriptor, Stockholm, 1977, 19-43.

Lancester, F.W.: "Toward Paperless Information Systems." Academic Press Inc., New York, 1978.

Lancaster, F.W.: "Information Retrieval Systems: Characteristics, Testing and Evaluation." Wiley, New York, 1979.

Lancaster, F.W.: "Trends in subject indexing from 1957 to 2000." Peter J. Taylor (ed.) New trends in documentation and information. Proceedings of the 39th FID Congress, Edinburgh, 1978. FID Publication 566. ASLIB. 1980, 223-233.

Langefors, B.: "Theoretical Analysis of Information Sysems." 3rd ed. Studentlitteratur, Lund, 1971.

Langefors, B. and Samuelson, K.: "Information and Data in Systems." Petrocelli/Charter, New York, 1976.

Lau, B. and Winkel, A: "Problems in Subject Description of Monographical Literature: Report on an Experiment". Int. Forum Info. and Doc, 8(1983)3, pp.8-12.

Lindsay, P.H. Norman, D.A.: "Human Information Processing." 2nd ed. Academic Press, New York, 1977.

Luhn, H.P.: "A Statistical Approach to Mechanised Encoding and Searching of Literary Information." IBM Journal of Research and Development, 1(1957)4, 309-317.

Luhn, H.P.: "The Automatic Creation of Literature Abstracts." IBM Journal of Research and Development, 2(1958)2, 159-165.

Luhn, H.P.: "Keyword-in-Context Index for Technical Literature (KWIC)." American Documentation, 11(1960)4, 288-295.

Luria, A.R.: "Cognitive development: its cultural and social foundations." Harvard University Press, London, 1976.

Mackay, E.M.: "Information, mechanism and meaning." M.I.T. Press, Cambridge, 1969.

Markey, K.: "Research Report on Analytical Review of Catalog Use Studies." OCLC, Dublin, OH, 1980. Available through ERIC as ED 186041.

Markey, K.: "Subject searching in library catalogs. Before and after the introduction of online catalogs." OCLC, Dublin, OH, 1984.

Maron, M.E.: "On indexing retrieval and the meaning of about." Journal of the ASIS. 28(1977)1.

McClure, C.R.: "Subject and added entries as access to information." Journal of Academic Librarianship. 2(1976)1, 9-14.

Miller, J.G.: "Living Systems." New York, McGraw-Hill, 1978.

Minsky, M.: "A framework for representing knowledge." P. Winstow (ed.). The psychology of computer vision. McGraw Hill, New York, 1975, 211-277.

Montgomery, C.: "Linguistics and Information Science." Journal of ASIS 23(1972)3, 195-219.

Moores, C. N.: "Analysis of the Hexagon of Notification." Journal of Documentation, 30(1974)2, 181-194.

Nauta, D. Jr.: "The Meaning of Information." The Hauge, Mouton, 1972.

Newell, A.: "Artificial intelligence and the concept of mind." R. Schank and K. M. Colby (eds.). Computer models of thought and language. San Francisco: Freeman, 1973.

Noel, J.: "Syntax, semantics and pragmatics in the automatic analysis of texts." Peter J. Taylor (ed.) New trends in documentation and information. Proceedings of the 39th FID Congress, Edinburgh, 1978. FID Pulbication 566. ASLIB 1980, 39-51.

O'Connel, J. D. et al: "Electronically expanding the citizen's world." IEEE Spectrum, 6(1969), 30-40.

Ogden, C. K. and Richards, I. A.: "The Meaning of Meaning." Routledge & Kegan Paul, London, 1949.

Olney, J. C.: "Library Cataloging and Classification." Report no. TM-1192, System Development Corp., Santa Monica, 1963.

"Online-emnesøgning." Rapport fra Arbejdsgruppen vedrørende online-emnesøgning vedsat af Bibliotekscentralen. København, Bibliotekscentralens Forlag, 1983.

Pejtersen, A.M.: "The meaning of 'about' in finction indexing and retrieval." Aslib Proceedings, 31(1979)5, 251-57.

Pejtersen, A.M.: "Design of a classification scheme for fiction based on an analysis of actual user-librarian communication, and use of the scheme for control of librarians' search strategies." O.Harbo and L. Kaiberg (eds.) Theory and Application of Information Research, Mansell, London, 1980, 146-159.

Piaget, J.: "Le comportement, moteur de l'évolution." Gallimard, Paris, 1976.

Pierce, J.C.: "Back-of-book subject indexing with APL: automated indexing for those without computer background." Information Processing & Management, Vol.14(1978), 85-91.

Pierce, J.R.: "Symbols, signals and noise: the nature and process of communication." Harper, New York, 1961.

Popper, K. R.: "Objective knowledge on evolutionary apparoach." Clarendon Press, Oxford, 1973. First published 1972.

Popper, K. R. and Eccles, J. C.: "The Self and its Brain." Springer, Berlin, 1977.

Pratt, A.D.: "The information of the image." Libri, 27(1977), 204-220.

Robertson, S.E.: "Theories and Models in Information Retrieval." Journal of Documentation, 33(1977) June, 126-148.

Robertson, S.E. and Sparck Jones, K.: "Relevance Weighting of Search Terms." Journal of ASIS, 23(1976)3, 129-146.

Sager, N.: "Perspective paper: Computational linguistics." D.E. Walker et al (eds.) "Natural language in Information Science." FID Publication 551, Skriptor, Stockholm, 1977, 75-100.

Salton, G.: "Automatic Information Organization and Retrieval." McGraw-Hill, New York, 1968.

Salton, G.: "Automatic Text Analysis." Science, (1970)168, 335-343.

Salton, G.: (ed.) "The SMART retrieval systems: experiments in automatic document processing." Prentice-Hall, Englewood Cliffs, N.J., 1971.

Salton, G.: "Proposals for Dynamic Library." L. Estabrook (ed.) Libraries in Postindustrial Society. Phoenix, Ariz., 1977, 273-305.

Salton, G. and McGill, M.: "Introduction to modern information; retrieval." McGraw-Hill, New York, 1983.

Samuelson, K.: "Free-text indexing to meet computers halfway." IEEE Transactions on Engineering Writing and Speech, Vol. EWS-13, (1970)2, September, 46-48.

Samuelson, K.: "Informatics by General Systems and Cybernetics." The Royal Institute of Technology and the University of Stockholm, Department of Information Processing and Computer Science, KTH, TRITA-IBADB-5011, Stockholm 1977.

Samuelson, K.: "Informatics with System Science and Cybernetics. - Conepts and Definitions." The Royal Institute of Technology and the University of Stockholm, Department of Information Processing and Compute Science, TRITA-IBADB-5012, Stockholm 1978.

Saussure, F: "Cours de linguistique générale." Publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger. 5.éd. (Bibliothèque scientifique.), Paris, 1960.

Schank, R. and Abelson, R.P.: "Scripts, plans, goals and understanding: an inquiry into human knowledge structures." Hillsdale, N.J.: Erlbaum, 1977.

Schank, R.C., Lehnert, W.: "Review of natural language processing." P. Wegner (ed.) Research Directions in Software Technology. MIT Press, Cambridge, Mass. 1979.

Schank, R.C., Lebowitz, M. and Birnbaum, L.: "An Integrated Understander". American Journal of Computational Linguistics, 6(1980)1, 13-30.

Schullian, D.M.: "Ancient Medieval and Renaissanse Libraries." Article on Libraries, Encyclopedia Amer. Vol. 17. The American Corp., New York, 1960 edition, 353-358.

Schumacher, H.H. et al: "The use of Selected Portions of Technical Documents as Sources of Index Terms and Effect on Input Costs and Retrieval Effectiveness." UDRI-TR-73-10, University of Dayton Research Institut, Ohio, 1973.

Settel, B.(ed.): "Subject description of books. A manual of procedures for augmenting subject description in library catalogs." Syracuse: Syracuse University, School of Information Studies. 1977.

Shannon, C.E.: "The Mathematical Theory of Communication." The University of Illinois Press, Urbana, 1949.

Shannon, C.E. and Weaver, W.: "The Mathematical Theory of Communication." The University of Illinois Press, Urbana, (1949)1969.

Shera, J.: "Of librarianship, documentation and information science." Unesco Bulletin for Libraries, 22(1968)2, 58-65.

Shera, J.: "Mechanization, librarianship, and the bibliographic enterprise." Journal of Documentation, 30(1974)2, 153-169.

Shera, J.H. and Cleveland, D.B.: "History and Foundations of Information Science." Williams M.E. (ed.). Annual Review of Information Science and Technology, Washington D.C., 12(1977), 249-275.

Shera, J. and Egan M.: "Foundations of a Theory of Bibliography." 1952.

Small, H: "Co-citation context analysis and the structure of paradigms." Journal of Documentation, 36(1980)3, 183-196.

Small, H. and Griffith, B.C: "The structure of scientific literature, I: Identifying and graphing specialties." Science Studies, 4(1974), 17-40.

Smith, L.C.: "Selected Artificial Intelligence Techniques in Information Retrieval Systems Research." Syracuse University, School of Information Studies, Ph.D. diss., Syracuse, 1979.

Smith, L.C.: "Artificial Intelligence Applications in Information Systems." Williams M.E. (ed.). Annual Review of Information Science and Technology, Washington D.C., 15(1980), 67-105.

Spang-Hanssen, H.: "Roles and Links compared with grammatical relations in natural languages." Dansk Teknisk Litteraturselskab Skriftserie, nr. 40. Lyngby, 1976.

Spang-Hanssen, H.: "The many virtues of the distinction between content and expression." Studia Linguistica XXXII. 1978, I-II.

Sparck Jones, K.: "Automatic Indexing 1974: A State of the Art Review." Computer Laboratory, Cambridge University, Cambridge. (1974a)

Sparck Jones, K.: "Progress in Documentation: Automatic Indexing." Journal of Documentation, 30(1974)4, 393-432. (1974b).

Sparck Jones, K.: "Automatic classification." A. Maltby (ed.), Classification in the 1970's. A second look. Clive Bingley, London, 1976, 209-225.

Sparck Jones, K.: "Problems in the representation of meaning in information retrieval." The Analysis of Meaning: INFORMATICS 5. Processing of ASLIB Conference 1979, 193-201.

Sparck Jones, K. and Kay, M.: "Linguistics and Information Science." New York and London, Academic Press, 1973.

Sparck Jones, K. and Kay, M.: "Linguistics and Information Science: After Five Years." D. E. Walker, et al (eds.) Natural Language in Information Science, FID Publication 551, Skriptor, Stockholm, 1977, 183-192.

Stevens, M. E.: "Automatic Indexing: A State-of-the-Art-Report." NBS Monograph 91. Washington D.C.: National Bureau of Standards. 1965. Reissued with additions in 1970.

Stokes, R.: "Bibliographical Control and Service." Andre Deutsch, London, 1965, p.30

Swift, D.F., et al: "'Aboutness' as a strategy for retrieval in the social sciences." ASLIB Proceedings, 39(1978)5, 182-187.

Tell, B.: "Document representation and indexer consistency." Proceedings of the ASIS, 6(1969), 285-291.

Tell, B.: "Titel, Titel och Nyckelord samt Referat som Sök-och Kostnadselement." En studie av POST-banden. Tidskrift för Dokumentation, 26(1970)5, 3-7.

Tell, B.: "Retrieval efficiency from titles and the cost of indexing." Information Storage and Retrieval, 7(1971), 241-243.

Tell, B.: "Review of 'Books are for use'." International Classification, 6(1979)1, Book Reviews, 47-48.

Thompson, J.: "Indexing creation on the Wang OIS 130." Electronic Library, 1(1983)3, 181-186.

Tulving, E. and Donaldson, W. (eds.): "Organization of memory." New York: Academic Press, 1972.

Walker, D. E. et al (eds.) "Nature Language in Information Science". FID Publication 551, Skriptor, Stockholm, 1977.

Wersig, G.: "Information-Kommunikation-Dokumentation." Pullach bei München, Verlag Dokumentation, 1971.

Wersig, G. and Neveling, U.: "The Phenomena of Interest to Information Science." The Information Scientist, 9(1975)4, 127-140.

Wheatley, H. B.: "What Is an Index?" Longmans, Green and Co., London, 1879.

White, I.: "We do, do, do, and don't know why. Catalog practices cry out for re-examination." Amercan Libraries, 11(1981)June, 317-318.

Wickelgren, W. A.: "Cognitive Psychology." Prentice Hall, Engelwood Cliffs, N.J. 1979.

Vickery, B. C.: "Information Systems." Butterworths, London, 1973.

Wilson, P.: "Two Kinds of Power: An Essay on Bibliographic Control." University of California Press, Berkeley, 1976.

Witty, F. J.: "The beginning of indexing and abstracting: Some notes towards a history of indexing and abstracting in antiquity and the Middle Ages." The Indexer, 8(1973)4, 193-198.

Wormell, I.: "Slagords-automatiserings-projektet - SAP." Tidskrift för Dokumentation, 34(1978)1, 3-6,18.

Wormell, I.: "Handbok för SAP-indexering." Lunds Universitetsbibliotek, Lund, 1979. LULID5/LIBR-1600.

Wormell, I.: "Slagords-automatiserings-projekt - SAP." Understött av SINFDOK och DFI, 1977-80. Interimsrapport. Lunds Universitetsbibliotek, Lund, 1981. LULID5/LIBR-1601.

Wormell, I.: "SOU-databasen: en ny typ av online katalog för faktasökning." Menneske-System-Maskin. Den 5. Nordiske IoD-konferanse, Trondheim, 1982, 317-325.

Wormell, I.: "SOU-databasen." Projekt vid Lunds Universitet 1980-83, understött av DFI. Slutrapport. Lunds Universitetsbibliotek, Lund, 1983. LULID/LIBR-1602.

Wormell, I.: "Indexes and text analysis in automated IR-systems". The Royal School of Librarianship, Copenhagen, 1984. (1984a).

Wormell, I.: "Cognitive aspects in nature language and free-text searching". Social Science Information Studies, 4(1984)4, 131-141. (1984b).

Vygotsky, L.S.: "Thought and language." Boston: MIT Press, 1962.

Zipf, G.K.: "Selected Studies of the Principle of Relative Frequencies of Language." Cambridge, Mass., 1932.

Zipf, G.K.: "Human Behaviour and the Principle of Least Effort". Addison Wesley Publishing, Reading, Massachusetts, 1949.

Zunde, P., Dexter, M.: "Indexing Consistency and Quality." American Documentation, 29(1969)3, 259-267.

Zunde, P.: "Empirial laws for information systems design". A. Debons and A. G. Larson (eds.) Information Science in Action: System Design. Proceedings of the NATO Advanced Study Institute of Information Science, Crete, 1978. Martinus Nijhoff Publishers, The Hague, 1983. pp. 152-171.